Union support recovery in high-dimensional multivariate regression

Guillaume Obozinski[†] Martin J. Wainwright[†,⋆] Michael I. Jordan[†,⋆]
{gobo, wainwrig, jordan}@stat.berkeley.edu

Department of Statistics[†], and Department of Electrical Engineering and Computer Science[⋆]
UC Berkeley, Berkeley, CA 94720

Technical Report, Department of Statistics, UC Berkeley
August 2008

Abstract

In the problem of multivariate regression, a K-dimensional response vector is regressed upon a common set of p covariates, with a matrix $B^* \in \mathbb{R}^{p \times K}$ of regression coefficients. We study the behavior of the group Lasso using ℓ_1/ℓ_2 regularization for the *union support problem*, meaning that the set of s rows for which B^* is non-zero is recovered exactly. Studying this problem under high-dimensional scaling, we show that group Lasso recovers the exact row pattern with high probability over the random design and noise for scalings of (n, p, s) such that the sample complexity parameter given by $\theta(n, p, s) := n/[2\psi(B^*)\log(p-s)]$ exceeds a critical threshold. Here n is the sample size, p is the ambient dimension of the regression model, s is the number of non-zero rows, and $\psi(B^*)$ is a *sparsity-overlap function* that measures a combination of the sparsities and overlaps of the K-regression coefficient vectors that constitute the model. This sparsity-overlap function reveals that, if the design is uncorrelated on the active rows, block ℓ_1/ℓ_2 regularization for multivariate regression never harms performance relative to an ordinary Lasso approach, and can yield substantial improvements in sample complexity (up to a factor of K) when the regression vectors are suitably orthogonal. For more general designs, it is possible for the ordinary Lasso to outperform the group Lasso. We complement our analysis with simulations that demonstrate the sharpness of our theoretical results, even for relatively small problems.

1 Introduction

The development of efficient algorithms for large-scale model selection has been a major goal of statistical learning research in the last decade. There is now a substantial body of work based on ℓ_1-regularization, dating back to the seminal work of Tibshirani (1996) and Donoho and collaborators (Chen et al., 1998; Donoho and Huo, 2001). The bulk of this work has focused on the standard problem of linear regression, in which one makes observations of the form

$$y \;=\; X\beta^* + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is a real-valued vector of observations, $w \in \mathbb{R}^n$ is an additive zero-mean noise vector, and $X \in \mathbb{R}^{n \times p}$ is the design matrix. A subset of the components of the unknown parameter vector $\beta^* \in \mathbb{R}^p$ are assumed non-zero; the model selection goal is to identify

these coefficients and (possibly) estimate their values. This goal can be formulated in terms of the solution of a penalized optimization problem:

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{n}\|y-X\beta\|_2^2+\lambda_n\|\beta\|_0\right\}, \tag{2}$$

where $\|\beta\|_0$ counts the number of non-zero components in β and where $\lambda_n > 0$ is a regularization parameter. Unfortunately, this optimization problem is computationally intractable, a fact which has led various authors to consider the convex relaxation (Tibshirani, 1996; Chen et al., 1998)

$$\arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{n}\|y-X\beta\|_2^2+\lambda_n\|\beta\|_1\right\}, \tag{3}$$

in which $\|\beta\|_0$ is replaced with the ℓ_1 norm $\|\beta\|_1$. This relaxation, often referred to as the Lasso (Tibshirani, 1996), is a quadratic program, and can be solved efficiently by various methods (e.g., Boyd and Vandenberghe, 2004; Osborne et al., 2000; Efron et al., 2004)).

A variety of theoretical results are now in place for the Lasso, both in the traditional setting where the sample size n tends to infinity with the problem size p fixed (Knight and Fu, 2000), as well as under high-dimensional scaling, in which p and n tend to infinity simultaneously, thereby allowing p to be comparable to or even larger than n (e.g., Meinshausen and Bühlmann, 2006; Wainwright, 2006; Zhao and Yu, 2006). In many applications, it is natural to impose *sparsity constraints* on the regression vector β^*, and a variety of such constraints have been considered. For example, one can consider a "hard sparsity" model in which β^* is assumed to contain at most s non-zero entries or a "soft sparsity" model in which β^* is assumed to belong to an ℓ_q ball with $q < 1$. Analyses also differ in terms of the loss functions that are considered. For the model or variable selection problem, it is natural to consider the $\{0-1\}$-loss associated with the problem of recovering the unknown support set of β^*. Alternatively, one can view the Lasso as a shrinkage estimator to be compared to traditional least squares or ridge regression; in this case, it is natural to study the ℓ_2-loss $\|\widehat{\beta}-\beta^*\|_2$ between the estimate $\widehat{\beta}$ and the ground truth. In other settings, the prediction error $\mathbb{E}[(Y-X^T\widehat{\beta})^2]$ may be of primary interest, and one tries to show risk consistency (namely, that the estimated model predicts as well as the best sparse model, whether or not the true model is sparse).

1.1 Block-structured regularization

While the assumption of sparsity at the level of individual coefficients is one way to give meaning to high-dimensional ($p \gg n$) regression, there are other structural assumptions that are natural in regression, and which may provide additional leverage. For instance, in a hierarchical regression model, groups of regression coefficients may be required to be zero or non-zero in a blockwise manner; for example, one might wish to include a particular covariate and all powers of that covariate as a group (Yuan and Lin, 2006; Zhao et al., 2007). Another example arises when we consider variable selection in the setting of multivariate regression: multiple regressions can be related by a (partially) shared sparsity pattern, such as when there are an underlying set of covariates that are "relevant" across regressions (Obozinski et al., 2007; Argyriou et al., 2006; Turlach et al., 2005; Zhang et al., 2008). Based on such motivations, a recent line of research (Bach et al., 2004; Tropp, 2006;

Yuan and Lin, 2006; Zhao et al., 2007; Obozinski et al., 2007; Ravikumar et al., 2008) has studied the use of *block-regularization schemes*, in which the ℓ_1 norm is composed with some other ℓ_q norm ($q > 1$), thereby obtaining the ℓ_1/ℓ_q norm defined as a sum of ℓ_q norms over groups of regression coefficients. The best known examples of such block norms are the ℓ_1/ℓ_∞ norm (Turlach et al., 2005; Zhang et al., 2008), and the ℓ_1/ℓ_2 norm (Obozinski et al., 2007).

In this paper, we investigate the use of ℓ_1/ℓ_2 block-regularization in the context of high-dimensional multivariate linear regression, in which a collection of K scalar outputs are regressed on the same design matrix $X \in \mathbb{R}^{n \times p}$. Representing the regression coefficients as an $p \times K$ matrix B^*, the multivariate regression model takes the form

$$Y \quad = \quad XB^* + W, \tag{4}$$

where $Y \in \mathbb{R}^{n \times K}$ and $W \in \mathbb{R}^{n \times K}$ are matrices of observations and zero-mean noise respectively. In addition, we assume a hard-sparsity model for the regression coefficients in which column j of the coefficient matrix B^* has non-zero entries on a subset

$$S_k := \{i \in \{1, \ldots, p\} \ \mid \ \beta^*_{ik} \neq 0\} \tag{5}$$

of size $s_k := |S_k|$. We focus on the problem of recovering the union of the supports, namely the set $S := \cup_{k=1}^K S_k$, corresponding to the subset of indices $i \in \{1, \ldots, p\}$ that are involved in at least one regression. This *union support problem* can be understood as the generalization of variable selection to the group setting. Rather than selecting specific components of a coefficient vector, we aim to select specific rows of a coefficient matrix. We thus also refer to the union support problem as the *row selection problem*. Note finally that recovering S is not equivalent to recovering each of the individual supports S_k.

If computational complexity were not a concern, the natural way to perform row selection for B^* would be by solving the optimization problem

$$\arg \min_{B \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2n} |\!|\!| Y - XB |\!|\!|_F^2 + \lambda_n \|B\|_{\ell_0/\ell_q} \right\}, \tag{6}$$

where $B = (\beta_{ik})_{1 \leq i \leq p, 1 \leq k \leq K}$ is a $p \times K$ matrix, the quantity $|\!|\!| \cdot |\!|\!|_F$ denotes the Frobenius norm[1], and the "norm" $\|B\|_{\ell_0/\ell_q}$ counts the number of rows in B that have non-zero ℓ_q norm. As before, the ℓ_0 component of this regularizer yields a non-convex and computationally intractable problem, so that it is natural to consider the relaxation

$$\arg \min_{B \in \mathbb{R}^{p \times K}} \left\{ \frac{1}{2n} |\!|\!| Y - XB |\!|\!|_F^2 + \lambda_n \|B\|_{\ell_1/\ell_q} \right\}, \tag{7}$$

where $\|B\|_{\ell_1/\ell_q}$ is the block ℓ_1/ℓ_q norm:

$$\|B\|_{\ell_1/\ell_q} \quad := \quad \sum_{i=1}^p \sqrt{\sum_{j=1}^K \beta_{ij}^q} = \sum_{i=1}^p \|\beta_i\|_q. \tag{8}$$

[1]The Frobenius norm of a matrix A is given by $|\!|\!| A |\!|\!|_F := \sqrt{\sum_{i,j} A_{ij}^2}$.

The relaxation (7) is a natural generalization of the Lasso; indeed, it specializes to the Lasso in the case $K = 1$. For later reference, we also note that setting $q = 1$ leads to the use of the ℓ_1/ℓ_1 block-norm in the relaxation (7). Since this norm decouples across both the rows and columns, this particular choice is equivalent to solving K separate Lasso problems, one for each column of the $p \times K$ regression matrix B^*. A more interesting choice is $q = 2$, which yields a block ℓ_1/ℓ_2 norm that couples together the columns of B. This regularization is commonly referred to as the *group Lasso*. As we discuss in Appendix 2, the group Lasso with $q = 2$ can be cast as a *second-order cone program* (SOCP), a family of optimization problems that can be solved efficiently with interior point methods (Boyd and Vandenberghe, 2004), and includes quadratic programs as a particular case.

Some recent work has addressed certain statistical aspects of block-regularization schemes. Meier et al. (2008) have performed an analysis of risk consistency with block-norm regularization. Bach (2008) provides an analysis of block-wise support recovery for the kernelized group-Lasso in the classical, fixed p setting. In the high-dimensional setting, Ravikumar et al. (2008) have studied the consistency of block-wise support recovery for the group-Lasso (ℓ_1/ℓ_2) for fixed design matrices, and their result is generalized by Liu and Zhang (2008) to block-wise support recovery in the setting of general ℓ_1/ℓ_q regularization, again for fixed design matrices. However, these analyses do not discriminate between various values of q, yielding the same qualitative results and the same convergence rates for $q = 1$ as for $q > 1$. Our focus, which is motivated by the empirical observation that the group Lasso can outperform the ordinary Lasso (Bach, 2008; Yuan and Lin, 2006; Zhao et al., 2007; Obozinski et al., 2007), is precisely the distinction between $q = 1$ and $q > 1$ (specifically $q = 2$).

The distinction between $q = 1$ and $q = 2$ is also significant from an optimization-theoretic point of view. In particular, the SOCP relaxations underlying the group Lasso ($q = 2$) are generally tighter than the quadratic programming relaxation underlying the Lasso ($q = 1$); however, the improved accuracy is generally obtained at a higher computational cost (Boyd and Vandenberghe, 2004). Thus we can view our problem as an instance of the general question of the relationship of statistical efficiency to computational efficiency: does the qualitatively greater amount of computational effort involved in solving the group Lasso always yield greater statistical efficiency? More specifically, can we give theoretical conditions under which solving the generalized Lasso problem (7) has greater statistical efficiency than naive strategies based on the ordinary Lasso? Conversely, can the group Lasso ever be worse than the ordinary Lasso?

With this motivation, this paper provides a detailed analysis of model selection consistency of the group Lasso (7) with ℓ_1/ℓ_2-regularization. Statistical efficiency is defined in terms of the scaling of the sample size n, as a function of the problem size p and sparsity structure of the regression matrix B^*, required for consistent row selection. Our analysis is high-dimensional in nature, allowing both n and p to diverge, and yielding explicit error bounds as a function of p. As detailed below, our analysis provides affirmative answers to both of the questions above. First, we demonstrate that under certain structural assumptions on the design and regression matrix B^*, the group ℓ_1/ℓ_2-Lasso is always guaranteed to out-perform the ordinary Lasso, in that it correctly performs row selection for sample sizes for which the Lasso fails with high probability. Second, we also exhibit some problems (though arguably not generic) for which the group Lasso will be outperformed by the naive strategy of applying the Lasso separately to each of the K columns, and taking the union

of supports.

1.2 Our results

The main contribution of this paper is to show that under certain technical conditions on the design and noise matrices, the model selection performance of block-regularized ℓ_1/ℓ_2 regression (7) is governed by the *sample complexity function*

$$\theta_{\ell_1/\ell_2}(n, p\,; B^*) \quad := \quad \frac{n}{2\,\psi(B^*)\log(p-s)}, \tag{9}$$

where n is the sample size, p is the ambient dimension, $s = |S|$ is the number of rows that are non-zero, and $\psi(\cdot)$ is a *sparsity-overlap function*. Our use of the term "sample complexity" for θ_{ℓ_1/ℓ_2} reflects the role it plays in our analysis as the rate at which the sample size must grow in order to obtain consistent row selection as a function of the problem parameters. More precisely, for scalings (n, p, s, B^*) such that $\theta_{\ell_1/\ell_2}(n, p\,; B^*)$ exceeds a fixed critical threshold $t^\star \in (0, +\infty)$, we show the probability of correct row selection by ℓ_1/ℓ_2 group Lasso converges to one.

Whereas the ratio $\frac{\log p}{n}$ is standard for high-dimensional theory on ℓ_1-regularization, the function $\psi(B^*)$ is a novel and interesting quantity, which measures both the sparsity of the matrix B^*, as well as the overlap between the different regression tasks, represented by the columns of B^*. (See equation (15) for the precise definition of $\psi(B^*)$.) As a particular illustration, consider the special case of a single-task or univariate regression with $K = 1$, in which the convex program (7) reduces to the ordinary Lasso (3). In this case, if the design matrix is drawn from the Standard Gaussian ensemble (i.e., $X_{ij} \sim N(0, 1)$, i.i.d), we show that the sparsity-overlap function reduces to $\psi(B^*) = s$, corresponding to the support size of the single coefficient vector. We thus recover as a corollary a previously known result (Wainwright, 2006): namely, the Lasso succeeds in performing exact support recovery once the ratio $n/[s\log(p-s)]$ exceeds a certain critical threshold. At the other extreme, for a genuinely multivariate problem with $K > 1$ and s non-zero rows, again for a Standard Gaussian design, when the regression matrix is "suitably orthonormal" relative to the design (see Section 2 for a precise definition), the sparsity-overlap function is given by $\psi(B^*) = s/K$. In this case, ℓ_1/ℓ_2 block-regularization has sample complexity lower by a factor of K relative to the naive approach of solving K separate Lasso problems. Of course, there is also a range of behavior between these two extremes, in which the gain in sample complexity varies smoothly as a function of the sparsity-overlap $\psi(B^*)$ in the interval $[\frac{s}{K}, s]$. On the other hand, we also show that for suitably correlated designs, it is possible that the sample complexity $\psi(B^*)$ associated with ℓ_1/ℓ_2 row selection is larger than that of the ordinary Lasso (ℓ_1/ℓ_1) approach.

The remainder of the paper is organized as follows. In Section 2, we provide a precise statement of our main result, discuss some of its consequences, and illustrate the close agreement between our theory and simulations. Section 3 is devoted to the proof of this main result, with the argument broken down into a series of steps. Technical results are deferred to the appendix. We conclude with a brief discussion in Section 4.

1.3 Notation

We collect here some notation used throughout the paper. For a (possibly random) matrix $M \in \mathbb{R}^{p \times K}$, we define the Frobenius norm $|||M|||_F := (\sum_{i,j} m_{ij}^2)^{1/2}$, and for parameters $1 \leq a \leq b \leq \infty$, the ℓ_a/ℓ_b block norm

$$\|M\|_{\ell_a/\ell_b} := \left\{ \sum_{i=1}^{p} \left(\sum_{k=1}^{K} |m_{ik}|^b \right)^{\frac{a}{b}} \right\}^{\frac{1}{a}}. \tag{10}$$

These vector norms on matrices should be distinguished from the (a, b)-operator norms

$$|||M|||_{a,b} := \sup_{\|x\|_b = 1} \|Mx\|_a, \tag{11}$$

although some norms belong to both families; see Lemma 5 in Appendix B. Important special cases of the latter include the spectral norm $|||M|||_{2,2}$ (also denoted $|||M|||_2$), and the ℓ_∞-operator norm $|||M|||_{\infty,\infty} = \max_{i=1,\ldots,p} \sum_{j=1}^{K} |M_{ij}|$, denoted $|||M|||_\infty$ for short.

2 Main result and some consequences

The analysis of this paper applies to random ensembles of multivariate linear regression problems, each of the form (4), where the noise matrix $W \in \mathbb{R}^{n \times K}$ is assumed to consist of i.i.d. elements $W_{ij} \sim N(0, \sigma^2)$. We consider random design matrices X with each row drawn in an i.i.d. manner from a zero-mean Gaussian $N(0, \Sigma)$, where $\Sigma \succ 0$ is a $p \times p$ covariance matrix. We note in passing that analogs of our results with different constants apply to any design with sub-Gaussian rows.[2] Although the block-regularized problem (7) need not have a unique solution in general, a consequence of our analysis is that in the regime of interest, the solution is unique, so that we may talk unambiguously about the estimated support $\widehat{S}$. The main object of study in this paper is the probability $\mathbb{P}[\widehat{S} = S]$, where the probability is taken both over the random choice of noise matrix W and random design matrix X. We study the behavior of this probability as elements of the triplet (n, p, s) tend to infinity.

2.1 Notation and assumptions

More precisely, our main result applies to sequences of models indexed by $(n, p(n), s(n))$, an associated sequence of $p \times p$ covariance matrices, and a sequence $\{B^*\}$ of coefficient matrices with row support

$$S := \{i \mid \beta_i^* \neq 0\} \tag{12}$$

of size $|S| = s = s(n)$. We use S^c to denote its complement (i.e., $S^c := \{1, \ldots, p\} \backslash S$). We let

$$b_{\min}^* := \min_{i \in S} \|\beta_i^*\|_2, \tag{13}$$

correspond to the minimal ℓ_2 row-norm of the coefficient matrix B^* over its non-zero rows.

We impose the following conditions on the covariance Σ of the design matrix:

[2] See Buldygin and Kozachenko (2000) for more details on sub-Gaussian random vectors.

(*A*1) **Bounded eigenspectrum:** There exists fixed constants $C_{\min} > 0$ and $C_{\max} < +\infty$ such that all eigenvalues of the $s \times s$ matrix Σ_{SS} are contained in the interval $[C_{\min}, C_{\max}]$.

(*A*2) **Mutual incoherence:** There exists a fixed incoherence parameter $\gamma \in (0, 1]$ such that

$$\left|\left|\left| \Sigma_{S^c S} (\Sigma_{SS})^{-1} \right|\right|\right|_{\infty} \quad \leq \quad 1 - \gamma.$$

(*A*3) **Self-incoherence:** There exist $D_{\max} < +\infty$ such that $\left|\left|\left| (\Sigma_{SS})^{-1} \right|\right|\right|_{\infty} \leq D_{\max}$.

Assumption *A*1 prevents excess dependence among elements of the design matrix associated with the support S; conditions of this form are required for model selection consistency or ℓ_2 consistency of the Lasso. The mutual incoherence assumption and self-incoherence assumptions also well known from previous work on variable selection consistency of the Lasso (Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2006). Although such incoherence assumptions are not needed in analyzing ℓ_2 or risk consistency, they are known to be necessary for model selection consistency of the Lasso. Indeed, in the absence of such conditions, it is always possible to make the Lasso fail, even with an arbitrarily large sample size. (However, see Meinshausen and Yu (2008) for methods that weaken the incoherence condition.) Note that these assumptions are trivially satisfied by the standard Gaussian ensemble $\Sigma = I_{p \times p}$, with $C_{\min} = C_{\max} = 1$, $D_{\max} = 1$, and $\gamma = 1$. More generally, it can be shown that various matrix classes (e.g., Toeplitz matrices, tree-structured covariance matrices, bounded off-diagonal matrices) satisfy these conditions (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2006).

2.2 Statement of main result

We require a few pieces of notation before stating the main result. For an arbitrary matrix $B_S \in \mathbb{R}^{s \times K}$ with i^{th} row $\beta_i \in \mathbb{R}^{1 \times K}$, we define the matrix $\zeta(B_S) \in \mathbb{R}^{s \times K}$ with i^{th} row

$$\zeta(\beta_i) \quad := \quad \frac{\beta_i}{\|\beta_i\|_2}. \tag{14}$$

With this notation, the *sparsity-overlap function* is given by

$$\psi(B) \quad := \quad \left|\left|\left| \zeta(B_S)^T (\Sigma_{SS})^{-1} \zeta(B_S) \right|\right|\right|_2, \tag{15}$$

where $|||\cdot|||_2$ denotes the spectral norm. Finally, the *sample complexity function* is given by

$$\theta_{\ell_1/\ell_2}(n, p\,; B^*) \quad := \quad \frac{n}{2\,\psi(B^*)\log(p - s)}. \tag{16}$$

With this setup, we have the following result:

Theorem 1. *Consider a random design matrix X drawn with i.i.d. $N(0, \Sigma)$ row vectors, where Σ satisfies assumptions A1 through A3, and an observation matrix Y specified by model (4). Suppose that the squared minimum value $(b^*_{\min})^2$ decays no more slowly than*

$f(p)\ \min\{\frac{1}{s},\ \frac{1}{\log(p-s)}\}$ *for some function* $f(p)/s \to 0$ *and* $f(p) \to +\infty$. *Then for all sequences* (n, p, B^*) *such that*

$$\theta_{\ell_1/\ell_2}(n, p; B^*) \ = \ \frac{n}{2\,\psi(B^*)\log(p-s)} \ > \ t^\star(\Sigma) \ := \ \frac{C_{\max}}{\gamma^2}, \tag{17}$$

we have with probability greater than $1 - c_1\exp(c_2 \log s)$*:*

(a) the SOCP (7) *with* $\lambda_n = \sqrt{\frac{f(p)\ \log p}{n}}$ *has a unique solution* $\widehat{B}$*, and*

(b) the row support set

$$\widehat{S} = S(\widehat{B}) \ := \ \Big\{ i \ \mid \ \widehat{\beta}_i \neq 0 \Big\} \tag{18}$$

specified by this unique solution is equal to the row support set $S(B^*)$ *of the true model.*

2.3 Some consequences of Theorem 1

We begin by making some simple observations about the sparsity overlap function.

Lemma 1. *(a) For any design satisfying assumption A*1*, the sparsity-overlap* $\psi(B^*)$ *obeys the bounds*

$$\frac{s}{C_{\max}K} \ \leq \ \psi(B^*) \ \leq \ \frac{s}{C_{\min}} \tag{19}$$

(b) If $\Sigma_{SS} = I_{s\times s}$*, and if the columns* $(Z^{(k)*})$ *of the matrix* $Z^* = \zeta(B^*)$ *are orthogonal, then the sparsity overlap function is* $\psi(B^*) = \max_k \|Z^{(k)*}\|^2$.

Proof. (a) To verify this claim, we first set $Z^*_S = \zeta(B^*_S)$, and use $Z^{(k)*}_S$ to denote the k^{th} column of Z^*_S. Since the spectral norm is upper bounded by the sum of eigenvalues, and lower bounded by the average eigenvalue, we have

$$\frac{1}{K}\text{tr}(Z^{*T}_S \Sigma^{-1}_{SS} Z^*_S) \leq \psi(B^*) \leq \text{tr}(Z^{*T}_S \Sigma^{-1}_{SS} Z^*_S).$$

Given our assumption $(A1)$ on Σ_{SS}, we have

$$\text{tr}(Z^{*T}_S \Sigma^{-1}_{SS} Z^*_S) = \sum_{k=1}^K Z^{(k)*T}_S \Sigma^{-1}_{SS} Z^{(k)*}_S \geq \frac{1}{C_{\max}} \sum_{k=1}^K \|Z^{(k)*}_S\|^2 \ = \ \frac{s}{C_{\max}},$$

using the fact that $\sum_{k=1}^K \|Z^{(k)*}_S\|^2 \ = \ \sum_{i=1}^s \|Z^*_i\|_2^2 = s$. Similarly, in the other direction, we have

$$\text{tr}(Z^{*T}_S \Sigma^{-1}_{SS} Z^*_S) = \sum_{k=1}^K Z^{(k)*T}_S \Sigma^{-1}_{SS} Z^{(k)*}_S \leq \frac{1}{C_{\min}} \sum_{k=1}^K \|Z^{(k)*}_S\|^2 \ = \ \frac{s}{C_{\min}},$$

which completes the proof.

(b) Under the assumed orthogonality, the matrix $Z^{*T}Z^*$ is diagonal with $\|Z^{(k)*}\|^2$ as the diagonal elements, so that the largest $\|Z^{(k)*}\|^2$ is then the largest eigenvalue of the matrix. □

Based on this lemma, we now study some special cases of Theorem 1. The simplest special case is the univariate regression problem ($K = 1$), in which case the quantity $\zeta(\beta^*)$ (as defined in equation (14)) simply outputs an s-dimensional sign vector with elements $z_i^* = \text{sign}(\beta_i^*)$. (Recall that the sign function is defined as $\text{sign}(0) = 0$, $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = -1$ if $x < 0$.) In this case, the sparsity overlap function is given by $\psi(\beta^*) = z^{*T}(\Sigma_{SS})^{-1}z^*$, and as a consequence of Lemma 1(a), we have $\psi(\beta^*) = \Theta(s)$. Consequently, a simple corollary of Theorem 1 is that the Lasso succeeds once the ratio $n/(2s\log(p-s))$ exceeds a certain critical threshold, determined by the eigenspectrum and incoherence properties of Σ. This result matches the necessary and sufficient conditions established in previous work on the Lasso (Wainwright, 2006).

We can also use Lemma 1 and Theorem 1 to compare the performance of the group Lasso to the following (arguably naive) strategy for row selection using the ordinary Lasso:

Row selection using ordinary Lasso:

1. Apply the ordinary Lasso separately to each of the K univariate regression problems specified by the columns of B^*, thereby obtaining estimates $\widehat{\beta}^{(k)}$ for $k = 1, \ldots, K$.
2. For $k = 1, \ldots, K$, estimate the column support via $\widehat{S}_k := \{i \mid \widehat{\beta}_i^{(k)} \neq 0\}$.
3. Estimate the row support by taking the union: $\widehat{S} = \cup_{k=1}^K \widehat{S}_k$.

To understand the conditions governing the success/failure of this procedure, note that it succeeds if and only if for each non-zero row $i \in S = \cup_{k=1}^K S_k$, the variable $\widehat{\beta}_i^{(k)}$ is non-zero for at least one k, and for all $j \in S^c = \{1, \ldots, p\} \backslash S$, the variable $\widehat{\beta}_j^{(k)} = 0$ for all $k = 1, \ldots, K$. From our understanding of the univariate case, we know that for $C = 2t^\star(\Sigma)$, the condition

$$n \;\geq\; C \max_{k=1,\ldots,K} \psi(\beta_S^{*(k)}) \log(p - s_k) \;\geq\; C \max_{k=1,\ldots,K} \psi(\beta_S^{*(k)}) \log(p-s) \tag{20}$$

is sufficient to ensure that the ordinary Lasso succeeds in row selection. Conversely, if $n < \max_{k=1,\ldots,K} \psi(\beta_S^{*(k)}) \log(p-s)$, then there will exist some $j \in S^c$ such for at least one $k \in \{1, \ldots, K\}$, there holds $\widehat{\beta}_j^{(k)} \neq 0$ with high probability, implying failure of the ordinary Lasso.

A natural question is whether the group Lasso, by taking into account the couplings across columns, always outperforms (or at least matches) this naive strategy. The following result shows that if the design is uncorrelated on its support, then indeed this is the case.

Corollary 1 (Group Lasso versus ordinary Lasso)**.** *Assume that $\Sigma_{SS} = I_{s\times s}$. Then for any multivariate regression problem, row selection using the ordinary Lasso strategy requires, with high probability, at least as many samples as the ℓ_1/ℓ_2 group Lasso. In particular, the relative efficiency of group Lasso versus ordinary Lasso is given by the ratio*

$$\frac{\max_{k=1,\ldots,K} \psi(\beta_S^{*(k)}) \log(p - s_k)}{\psi(B_S^*) \log(p-s)} \quad \geq \quad 1. \tag{21}$$

Proof. From our discussion preceding the statement of Corollary 1, we know that the quantity

$$\max_{k=1,\ldots,K} \psi(\beta_S^{*(k)}) \log(p - s_k) \;=\; \max_{k=1,\ldots,K} s_k \log(p - s_k) \;\geq\; \max_{k=1,\ldots,K} s_k \log(p - s)$$

governs the performance of the ordinary Lasso procedure for row selection. It remains to show then that $\psi(B_S^*) \leq \max_k s_k$.

As before, we use the notation $Z_S^* = \zeta(B_S^*)$, and Z_i^* for the i^{th} row of Z_S^*. Since $\Sigma_{SS} = I_{s \times s}$, we have $\psi(B^*) = \|Z_S^*\|_2^2$. Consequently, by the variational representation of the ℓ_2-norm, we have

$$\psi(B^*) \;=\; \max_{x \in \mathbb{R}^K \,:\, \|x\| \leq 1} \|Z_S^* x\|^2 \leq \max_{x \in \mathbb{R}^K \,:\, \|x\| \leq 1} \sum_{i=1}^{s} \left(Z_i^{*T} x \right)^2.$$

Let $|Z_i^*| = (|Z_{i1}^*|, \ldots, |Z_{ik}^*|)^T$ and $y_i = (x_1 \operatorname{sign}(Z_{i1}^*), \ldots, x_K \operatorname{sign}(Z_{iK}^*))^T$. By the Cauchy-Schwartz inequality,

$$\left(Z_i^{*T} x \right)^2 = \left(|Z_i^*|^T y_i \right)^2 \leq \| \, |Z_i^*| \, \|^2 \|y_i\|^2 = \|Z_i^*\|^2 \sum_k x_k^2 \operatorname{sign}(Z_{ik}^*)^2$$

so that, if $\|x\| \leq 1$, we have

$$\sum_{i=1}^{s} \left(Z_i^{*T} x \right)^2 \leq \sum_{i=1}^{s} \|Z_i^*\|^2 \sum_{k=1}^{K} x_k^2 \operatorname{sign}(Z_{ik}^*)^2 = \sum_{k=1}^{K} x_k^2 \sum_{i=1}^{s} \operatorname{sign}(Z_{ik}^*)^2 = \sum_{k=1}^{K} x_k^2 s_k \leq \max_{1 \leq k \leq K} s_k,$$

thereby establishing the claim. □

We illustrate Corollary 1 by considering some special cases:

Example 1 (Identical regressions). Suppose that $B^* := \beta^* \vec{1}_K^T$—that is, B^* consists of K copies of the same coefficient vector $\beta^* \in \mathbb{R}^p$, with support of cardinality $|S| = s$. We then have $[\zeta(B^*)]_{ij} = \operatorname{sign}(\beta_i^*)/\sqrt{K}$, from which we see that $\psi(B^*) = z^{*T}(\Sigma_{SS})^{-1} z^*$, with z^* being an s-dimensional sign vector with elements $z_i^* = \operatorname{sign}(\beta_i^*)$. Consequently, we have the equality $\psi(B^*) = \psi(\beta^{(1)*})$, so that there is no benefit in using the group Lasso relative to the strategy of solving separate Lasso problems and constructing the union of individually estimated supports. This fact might seem rather pessimistic, since under model (4), we essentially have Kn observations of the coefficient vector β^* with the same design matrix but K independent noise realizations. However, under the given conditions, the rates of convergence for model selection in high-dimensional results such as Theorem 1 are determined by the number of interfering variables, $p - s$, as opposed to the noise variance.

In contrast to this pessimistic example, we now turn to the most optimistic extreme:

Example 2 ("Orthonormal" regressions). Suppose that $(\Sigma_{SS}) = I_{s \times s}$ and (for $s > K$) suppose that B^* is constructed such that the columns of the $s \times K$ matrix $\zeta(B^*)$ are all

orthonormal. Under these conditions, we claim that the sample complexity of group Lasso is lower than that of the ordinary Lasso by a factor of $1/K$. Indeed, we observe that

$$K\psi(B^*) = K\|Z^{(1)*}\|^2 = \sum_{k=1}^{K} \|Z^{(k)*}\|^2 = \mathrm{tr}\left(Z^{*T}Z^*\right) = \mathrm{tr}\left(Z^*Z^{*T}\right) = s,$$

because $Z^*Z^{*T} \in \mathbb{R}^{s\times s}$ is the Gram matrix of s unit vectors of $\mathbb{R}^k$ and its diagonal elements are therefore all equal to 1. Consequently, the group Lasso recovers the row support with high probability for sequences such that

$$\frac{n}{2\,\frac{s}{K}\,\log(p-s)} \quad > \quad 1,$$

which allows for sample sizes $1/K$ smaller than the ordinary Lasso approach.

Corollary 1 and the subsequent examples address the case of uncorrelated design ($\Sigma_{SS} = I_{s\times s}$) on the row support S, for which the group Lasso is never worse than the ordinary Lasso in performing row selection. The following example shows that if the supports are disjoint, the ordinary Lasso has the same sample complexity as the group Lasso for uncorrelated design $\Sigma_{SS} = I_{s\times s}$, but can be better than the group Lasso for designs Σ_{SS} with suitable correlations:

Corollary 2 (Disjoint supports)**.** *Suppose that the support sets S_k of individual regression problems are all disjoint. Then for any design covariance Σ_{SS}, we have*

$$\max_{k=1,\ldots,K} \psi(\beta^{(k)*}) \quad \overset{(a)}{\leq} \quad \psi(B^*) \quad \overset{(b)}{\leq} \quad \sum_{k=1}^{K} \psi(\beta^{(k)*}) \tag{22}$$

Proof. First note that, since all supports are disjoint, $Z_i^{(k)*} = \mathrm{sign}(\beta_{ik}^*)$, so that $Z_S^{(k)*} = \zeta(\beta_S^{(k)*})$. Inequality (b) is then immediate, since $\left|\!\left|\!\left| Z_S^{*T}\Sigma_{SS}^{-1}Z_S^* \right|\!\right|\!\right|_2 \leq \mathrm{tr}(Z_S^{*T}\Sigma_{SS}^{-1}Z_S^*)$. To establish inequality (a), we note that

$$\psi(B^*) = \max_{x\in\mathbb{R}^K\,:\,\|x\|\leq 1} x^T Z_S^{*T}\Sigma_{SS}^{-1}Z_S^* x \geq \max_{1\leq k\leq K} e_k^T Z_S^{*T}\Sigma_{SS}^{-1}Z_S^* e_k = \max_{1\leq k\leq K} Z_S^{(k)*T}\Sigma_{SS}^{-1}Z_S^{(k)*}.$$

□

We illustrate Corollary 2 with an example.

Example 3. Disjoint support with uncorrelated design Suppose that $\Sigma_{SS} = I_{s\times s}$, and the supports are disjoint. In this case, we claim that the sample complexity of the ℓ_1/ℓ_2 group Lasso is the same as the ordinary Lasso. If the individual regressions have disjoint support, then $Z_S^* = \zeta(B_S^*)$ has only a single non-zero entry per row and therefore the columns of Z^* are orthogonal. Moreover, $Z_{ik}^* = \mathrm{sign}(\beta_i^{(k)*})$. By Lemma 1(b), the sparsity-overlap function $\psi(B^*)$ is equal to the largest squared column norm. But $\|Z^{(k)*}\|^2 = \sum_{i=1}^{s} \mathrm{sign}(\beta_i^{(k)*})^2 = s_k$. Thus, the sample complexity of the group Lasso is the same as the ordinary Lasso in this case.[3]

[3]In making this assertion, we are ignoring any difference between $\log(p-s_k)$ and $\log(p-s)$, which is valid, for instance, in the regime of sublinear sparsity, when $s_k/p \to 0$ and $s/p \to 0$.

Finally, we consider an example that illustrates the effect of correlated designs:

Example 4. Effects of correlated designs To illustrate the behavior of the sparsity-overlap function in the presence of correlations in the design. we consider the simple case of two regressions with support of size 2. For parameters ϑ_1 and $\vartheta_2 \in [0, \pi]$ and $\rho \in (-1, +1)$, consider regression matrices B^* such that $B^* = \zeta(B^*_S)$ and

$$\zeta(B^*_S) = \begin{bmatrix} \cos(\vartheta_1) & \sin(\vartheta_1) \\ \cos(\vartheta_2) & \sin(\vartheta_2) \end{bmatrix} \qquad \text{and} \qquad \Sigma^{-1}_{SS} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{23}$$

Setting $M^* = \zeta(B^*_S)^T \Sigma^{-1}_{SS} \zeta(B^*_S)$, a simple calculation shows that

$$\operatorname{tr}(M^*) = 2(1 + \rho \cos(\vartheta_1 - \vartheta_2)), \qquad \text{and} \qquad \det(M^*) = (1 - \rho^2) \sin(\vartheta_1 - \vartheta_2)^2,$$

so that the eigenvalues of M^* are

$$\mu^+ = (1 + \rho)(1 + \cos(\vartheta_1 - \vartheta_2)), \qquad \text{and} \qquad \mu^- = (1 - \rho)(1 - \cos(\vartheta_1 - \vartheta_2)).$$

so that $\psi(B^*) = \max(\mu^+, \mu^-)$. On the other hand, with

$$\tilde{z_1} = \zeta(\beta^{(1)*}) = \begin{pmatrix} \operatorname{sign}(\cos(\vartheta_1)) \\ \operatorname{sign}(\cos(\vartheta_2)) \end{pmatrix} \quad \text{and} \quad \tilde{z}_2 = \zeta(\beta^{(2)*}) = \begin{pmatrix} \operatorname{sign}(\sin(\vartheta_1)) \\ \operatorname{sign}(\sin(\vartheta_2)) \end{pmatrix}, \quad \text{we have}$$

$$\psi(\beta^{(1)*}) = \tilde{z}_1^T \Sigma^{-1}_{SS} \tilde{z}_1 = \mathbf{1}_{\{\cos(\vartheta_1) \neq 0\}} + \mathbf{1}_{\{\cos(\vartheta_2) \neq 0\}} + 2\,\rho\, \operatorname{sign}(\cos(\vartheta_1)\cos(\vartheta_2)),$$
$$\psi(\beta^{(2)*}) = \tilde{z}_2^T \Sigma^{-1}_{SS} \tilde{z}_2 = \mathbf{1}_{\{\sin(\vartheta_1) \neq 0\}} + \mathbf{1}_{\{\sin(\vartheta_2) \neq 0\}} + 2\,\rho\, \operatorname{sign}(\sin(\vartheta_1)\sin(\vartheta_2)).$$

Figure 4 provides a graphical comparison of these sample complexity functions. The function $\tilde{\psi}(B^*) = \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ is discontinuous on $\mathcal{S} = \frac{\pi}{2}\mathbb{Z} \times \mathbb{R} \cup \mathbb{R} \times \frac{\pi}{2}\mathbb{Z}$, and, as a consequence, so is its difference with $\psi(B^*)$. Note that, for fixed ϑ_1 or fixed ϑ_2, some of these discontinuities are *removable discontinuities* of the induced function on the other variable, and these discontinuities therefore create needles, slits or flaps in the graph of the function $\tilde{\psi}$. Denote by $\mathcal{R}^+$ (resp. $\mathcal{R}^-$) the set $\mathcal{R}^+ = \{(\vartheta_1, \vartheta_2) | \min[\cos(\vartheta_1)\cos(\vartheta_2), \sin(\vartheta_1)\sin(\vartheta_2)] > 0\}$, (resp. $\mathcal{R}^- = \{(\vartheta_1, \vartheta_2) | \max[\cos(\vartheta_1)\cos(\vartheta_2), \sin(\vartheta_1)\sin(\vartheta_2)] < 0\}$) on which $\tilde{\psi}(B^*)$ reaches its minimum value when $\rho \geq 0.5$ (resp. when $\rho \leq 0.5$) (see middle and bottom center plots in figure 4). For $\rho = 0$, the top center graph illustrates that $\tilde{\psi}(B^*)$ is equal to 2 except for the cases of matrices B^*_S with disjoint support, corresponding to the discrete set $\mathcal{D} = \{(k\frac{\pi}{2}, (k \pm 1)\frac{\pi}{2}), k \in \mathbb{Z}\}$ for which it equals 1. The top rightmost graph illustrates that, as shown in Corollary 1, the inequality always holds for an uncorrelated design. For $\rho > 0$, the inequality $\psi(B^*) \leq \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ is violated only on a subset of $\mathcal{S} \cup \mathcal{R}^-$; and for $\rho < 0$, the inequality is symmetrically violated on a subset of $\mathcal{S} \cup \mathcal{R}^+$ (see Fig. 4).

2.4 Illustrative simulations

In this section, we provide the results of some simulations to illustrate the sharpness of Theorem 1, and furthermore to ascertain how quickly the predicted behavior is observed as elements of the triple (n, p, s) grow in different regimes. We explore the case of two regression tasks (i.e., $K = 2$) which share an identical support set S with cardinality $|S| = s$ in Section 2.4.1 and consider a slightly more general case in Section 2.4.2.

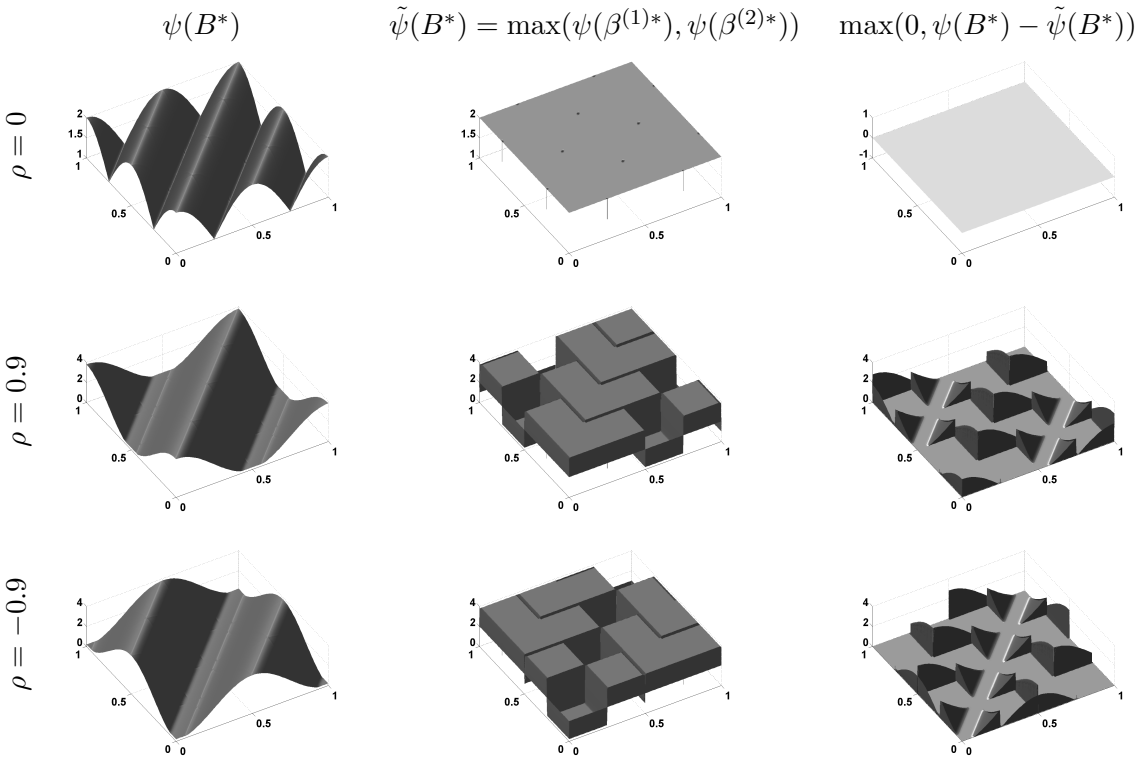

Figure 1. Comparison of sparsity-overlap functions for ℓ_1/ℓ_2 and the Lasso. For the pair $\frac{1}{2\pi}(\vartheta_1, \vartheta_2)$, we represent in each row of plots, corresponding respectively to $\rho = 0$ (top), 0.9 (middle) and -0.9 (bottom), from left to right, the quantities: $\psi(B^*)$ (left), $\max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ (center) and $\max(0, \psi(B^*) - \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*})))$ (right). The latter indicates when the inequality $\psi(B^*) \leq \max(\psi(\beta^{(1)*}), \psi(\beta^{(2)*}))$ does not hold and by how much it is violated.

2.4.1 Phase transition behavior

This first set of experiments is designed to reveal the phase transition behavior predicted by Theorem 1. The design matrix X is sampled from the standard Gaussian ensemble, with i.i.d. entries $X_{ij} \sim N(0, 1)$. We consider two types of sparsity,

- logarithmic sparsity, where $s = \alpha \log(p)$, for $\alpha = 2/\log(2)$, and
- linear sparsity, where $s = \alpha p$, for $\alpha = 1/8$,

for various ambient model dimensions $p \in \{16, 32, 64, 256, 512, 1024\}$. For a given triplet (n, p, s), we solve the block-regularized problem (7) with the regularization parameter $\lambda_n = \sqrt{\log(p-s)\,(\log s)/n}$. For each fixed (p, s) pair, we measure the sample complexity in terms of a parameter θ, in particular letting $n = \theta s \log(p - s)$ for $\theta \in [0.25, 1.5]$.

We let the matrix $B^* \in \mathbb{R}^{p \times 2}$ of regression coefficients have entries β^*_{ij} in $\{-1/\sqrt{2}, 1/\sqrt{2}\}$, choosing the parameters to vary the angle between the two columns, thereby obtaining various desired values of $\psi(B^*)$. Since $\Sigma = I_{p \times p}$ for the standard Gaussian ensemble, the sparsity-overlap function $\psi(B^*)$ is simply the maximal eigenvalue of the Gram matrix

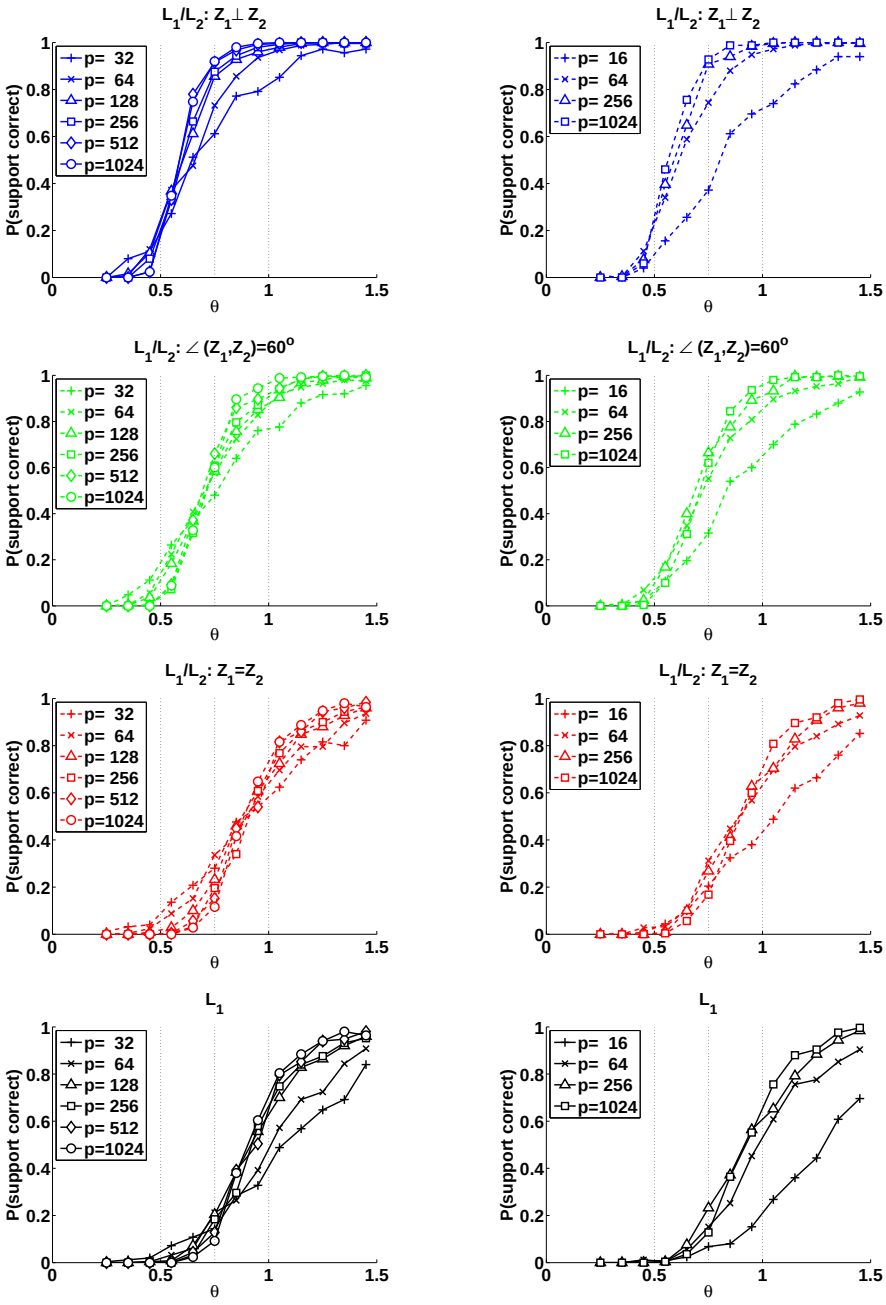

Figure 2. Plots of support union recovery probability $\mathbb{P}[\widehat{S} = S]$ versus the control parameter $\theta = n/[2s\log(p-s)]$ for two different types of sparsity, linear sparsity in the left column ($s = p/8$) and logarithmic sparsity in the right column ($s = 2\log_2(p)$)) and using ℓ_1/ℓ_2 regularization in the three first rows to estimate the support respectively in the three cases of identical regression, intermediate angles and orthonormal regressions. The fourth row presents results for the Lasso in the case of identical regressions.

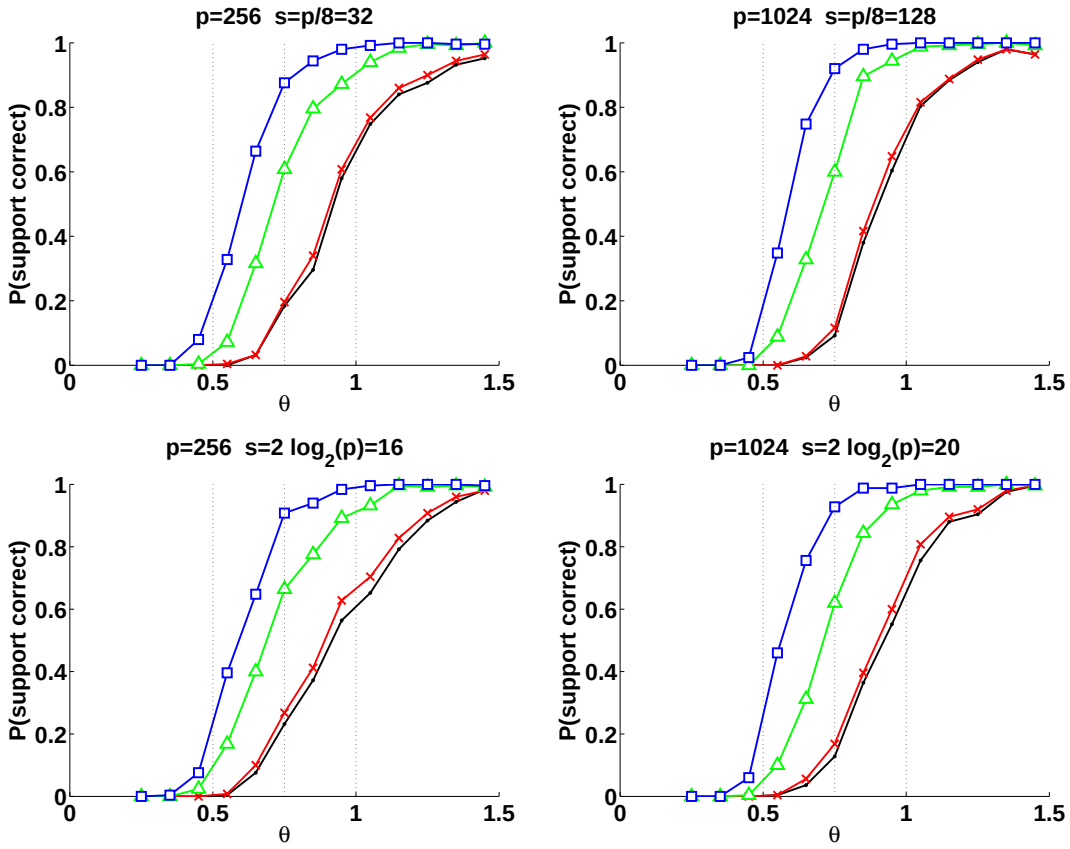

Figure 3. Plots of support recovery probability $\mathbb{P}[\widehat{S}=S]$ versus the control parameter $\theta = n/[2s\log(p-s)]$ for two different type of sparsity, logarithmic sparsity on top ($s = \mathcal{O}(\log(p))$) and linear sparsity on bottom ($s = \alpha p$), and for increasing values of p from left to right. The noise level is set at $\sigma = 0.1$. Each graph shows four curves (black,red, green,blue) corresponding to the case of independent ℓ_1 regularization, and, for ℓ_1/ℓ_2 regularization, the cases of identical regression, intermediate angles, and "orthonormal" regressions. Note how curves corresponding to the same case across different problem sizes p all coincide, as predicted by Theorem 1. Moreover, consistent with the theory, the curves for the identical regression group reach $\mathbb{P}[\widehat{S}=S] \approx 0.50$ at $\theta \approx 1$, whereas the orthogonal regression group reaches 50% success substantially earlier.

$\zeta(B_S^*)^T\zeta(B_S^*)$. Since $|\beta_{ij}^*| = 1/\sqrt{2}$ by construction, we are guaranteed that $B_S^* = \zeta(B_S^*)$, that the minimum value $b_{\min}^* = 1$, and moreover that the columns of $\zeta(B_S^*)$ have the same Euclidean norm.

To construct parameter matrices B^* that satisfy $|\beta_{ij}| = 1/\sqrt{2}$, we choose both p and the sparsity scalings so that the obtained values for s are multiples of four. We then construct the columns $Z^{(1)*}$ and $Z^{(2)*}$ of the matrix $B^* = \zeta(B^*)$ from copies of vectors of length four. Denoting by $\otimes$ the usual matrix tensor product, we consider the following 4-vectors:

Identical regressions: We set $Z^{(1)*} = Z^{(2)*} = \frac{1}{\sqrt{2}}\vec{1}_s$, so that the sparsity-overlap function is $\psi(B^*) = s$.

Orthonormal regressions: Here B^* is constructed with $Z^{(1)*} \perp Z^{(2)*}$, so that $\psi(B^*) = \frac{s}{2}$, the most favorable situation. In order to achieve this orthonormality, we set $Z^{(1)*} = \frac{1}{\sqrt{2}}\vec{1}_s$ and $Z^{(2)*} = \frac{1}{\sqrt{2}}\vec{1}_{s/2} \otimes (1, -1)^T$.

Intermediate angles: In this intermediate case, the columns $Z^{(1)*}$ and $Z^{(2)*}$ are at a 60° angle, which leads to $\psi(B^*) = \frac{3}{4}s$. Specifically, we set $Z^{(1)*} = \frac{1}{\sqrt{2}}\vec{1}_s$ and $Z^{(2)*} = \frac{1}{\sqrt{2}}\vec{1}_{s/4} \otimes (1, 1, 1, -1)^T$.

Figures 2 shows plots of linear sparsity (left column) and logarithmic sparsity (right column) for all three cases solved using the group ℓ_1/ℓ_2 relaxation (top three rows), as well as the reference Lasso case for the case of identical regressions (bottom row). Each panel plots the success probability $\mathbb{P}[\widehat{S} = S]$ versus the rescaled sample size $\theta = n/[2s\log(p-s)]$. Under this re-scaling, Theorem 1 predicts that the curves should align, and that the success probability should transition to 1 once θ exceeds a critical threshold (dependent on the type of ensemble). Note that for suitably large problem sizes ($p \geq 128$), the curves do align in the predicted way, showing step-function behavior. Figure 3 plots data from the same simulations in a different format. Here the top row corresponds to logarithmic sparsity, and the bottow row to linear sparsity; each panel shows the four different choices for B^*, with the problem size p increasing from left to right. Note how in each panel the location of the transition of $\mathbb{P}[\widehat{S} = S]$ to one shifts from right to left, as we move from the case of identical regressions to intermediate angles to orthogonal regressions.

2.4.2 Empirical thresholds

In this experiment, we aim at verifying more precisely the location of the ℓ_1/ℓ_2 threshold as the regression vectors vary continuously from identical to orthonormal. We consider the case of matrices B^* of size $s \times 2$ for s even. In Example Sec. 4 of Sec. 2.3, we characterized the value of $\psi(B^*)$ if B^* is a 2×2 matrix.

In order to generate a family of regression matrices with smoothly varying sparsity/overlap function consider the following 2×2 matrix:

$$B_1(\alpha) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \cos(\frac{\pi}{4} + \alpha) & \sin(\frac{\pi}{4} + \alpha) \end{bmatrix}. \tag{24}$$

Note that α is the angle between the two *rows* of $B_1(\alpha)$ in this setup. Note moreover that the columns of $B_1(\alpha)$ have varying norm.

We use this base matrix to define the following family of regression matrices $B_S^* \in \mathbb{R}^{s \times 2}$:

$$\mathcal{B}_1 \quad := \quad \left\{ B_{1s}(\alpha) = \vec{1}_{s/2} \otimes B_1(\alpha),\ \alpha \in \left[0, \frac{\pi}{2}\right] \right\}. \tag{25}$$

For a design matrix drawn from the Standard Gaussian ensemble, the analysis of Example Sec. 4 in Sec. 2.3 naturally extends to show that the sparsity/overlap function is $\psi(B_{s1}(\alpha)) = \frac{s}{2}(1 + |\cos(\alpha)|)$. Moreover, as we vary α from 0 to $\frac{\pi}{2}$, the two regressions vary from identical to "orthonormal" and the sparsity/overlap function decreases from s to $\frac{s}{2}$.

We fix the problem size $p = 2048$ and sparsity $s = \log_2(p) = 22$. For each value of $\alpha \in [0, \frac{\pi}{2}]$, we generate a matrix from the specified family and angle. We then solve

the block-regularized problem (7) with sample size $n = 2\theta s \log(p-s)$ for a range of θ in $[.25, 1.5]$; for each value of θ, we repeat the experiment (generating random design matrix X and observation matrix W each time) over $T = 500$ trials. Based on these trials, we then estimate the value of $\theta_{50\%}$ for which the exact support is retrieved at least 50% of the time. Since $\psi(B^*) = \frac{1+|\cos(\alpha)|}{2} s$, our theory predicts that if we plot $\theta_{50\%}$ versus $|\cos(\alpha)|$, it should lie on or below the straight line $\frac{1+|\cos(\alpha)|}{2}$. We also perform the same experiments for row selection using the ordinary Lasso, and plot the resulting estimated thresholds on the same axes.

The results are shown in Figure 4. Note first that the curve obtained for $\widehat{S}_{\ell_1/\ell_2}$ (blue circles) coincides roughly with the theoretical prediction $\frac{1+|\cos(\alpha)|}{2}$ (black dashed diagonal) as regressions vary from orthogonal to identical. Moreover, the estimated $\theta_{50\%}$ of the ordinary Lasso remains above 0.9 for all values of α, which is close to the theoretical value of 1. However, the curve obtained is not constant, but is roughly sigmoidal with a first plateau close to 1 for $\cos(\alpha) < 0.4$ and a second plateau close to 0.9 for $\cos(\alpha) > 0.5$. The latter coincides with the empirical value of $\theta_{50\%}$ for the univariate Lasso for the first column $\beta^{(1)*}$ (not shown). There are two reasons why the value of $\theta_{50\%}$ for the ordinary Lasso does not match the prediction of the first-order asymptotics: first, for $\alpha = \frac{\pi}{4}$ (corresponding to $\cos(\alpha) = 0.7$), the support of $\beta^{(2)*}$ is reduced by one half and therefore its sample complexity is decreased in that region. Second, the supports recovered by individual Lassos for $\beta^{(1)*}$ and $\beta^{(2)*}$ vary from uncorrelated when $\alpha = \frac{\pi}{2}$ to identical when $\alpha = 0$. It is therefore not surprising that the sample complexity is the same as a single univariate Lasso for $\cos(\alpha)$ large and higher for $\cos(\alpha)$ small, where independent estimates of the support are more likely to include, by union, spurious covariates in the row support.

3 Proof of Theorem 1

In this section, we provide the proof of our main result. For the convenience of the reader, we begin by recapitulating the notation to be used throughout the argument.

- the sets S and S^c are a partition of the set of columns of X, such that $|S| = s$, $|S^c| = p - s$, and
- the design matrix is partitioned as $X = \begin{bmatrix} X_S & X_{S^c} \end{bmatrix}$, where $X_S \in \mathbb{R}^{n \times s}$ and $X_{S^c} \in \mathbb{R}^{n \times (p-s)}$.
- the regression coefficient matrix is also partitioned as $B^* = \begin{bmatrix} B^*_S \\ B^*_{S^c} \end{bmatrix}$, with $B^*_S \in \mathbb{R}^{s \times K}$ and $B^*_{S^c} = 0 \in \mathbb{R}^{(p-s) \times K}$. We use β^*_i to denote the i^{th} row of B^*.
- the regression model is given by $Y = XB^* + W$, where the noise matrix $W \in \mathbb{R}^{n \times K}$ has i.i.d. $N(0, \sigma^2)$ entries.
- The matrix $Z^*_S = \zeta(B^*_S) \in \mathbb{R}^{s \times K}$ has rows $Z^*_i = \zeta(\beta^*_i) = \frac{\beta^*_i}{\|\beta^*_i\|_2} \in \mathbb{R}^K$.

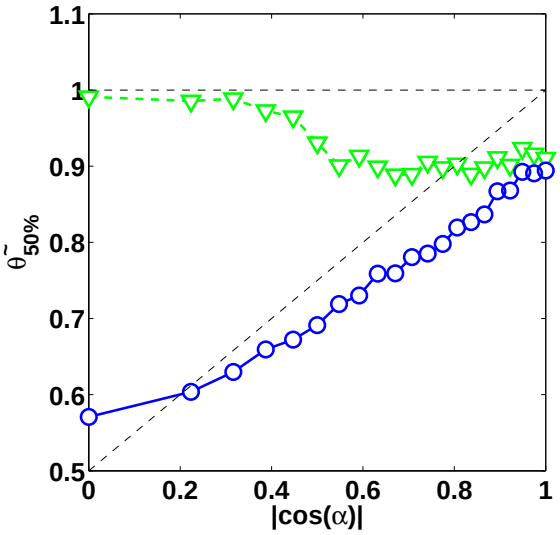

Figure 4. Plots of the Lasso sample complexity $\theta = n/[2s\log(p-s)]$ for which the probability of union support recovery exceeds 50% empirically as a function of $|\cos(\alpha)|$ for ℓ_1-based recovery and ℓ_1/ℓ_2 based recovery, where α is the angle between $Z^{(1)*}$ and $Z^{(2)*}$ for the family $\mathcal{B}_1$. We consider the two following methods for performing row selection: Ordinary Lasso (ℓ_1, green triangles) and group ℓ_1/ℓ_2 Lasso (blue circles).

3.1 High-level proof outline

At a high level, the proof is based on the notion of a *primal-dual witness*: we construct a primal matrix $\widehat{B}$ along with a dual matrix $\widehat{Z}$ such that:

(a) the pair $(\widehat{B}, \widehat{Z})$ together satisfy the Karush-Kuhn-Tucker (KKT) conditions associated with the second-order cone program (7), and

(b) this solution certifies that the SOCP recovers the union of supports S.

For general high-dimensional problems (with $p \gg n$), the SOCP (7) need not have a unique solution; however, a consequence of our theory is that the constructed solution $\widehat{B}$ is the unique optimal solution under the conditions of Theorem 1.

We begin by noting that the block-regularized problem (7) is convex, and not differentiable for all B. In particular, denoting by β_i the i^{th} row of B, the subdifferential of the norm ℓ_1/ℓ_2-block norm over row i takes the form

$$[\partial \|B\|_{\ell_1/\ell_2}]_i \;=\; \begin{cases} \frac{\beta_i}{\|\beta_i\|_2} & \text{if } \beta_i \neq \vec{0} \\ Z_i \quad \text{such that } \|Z_i\|_2 \leq 1 & \text{otherwise.} \end{cases} \tag{26}$$

We also use the shorthand $\zeta(B_i) = \beta_i/\|\beta_i\|_2$ with an analogous definition for the matrix $\zeta(B_S)$, assuming that no row of B_S is identically zero. In addition, we define the *empirical covariance matrix*

$$\widehat{\Sigma} \;:=\; \frac{1}{n}X^T X \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T, \tag{27}$$

where X_i is the i^{th} column of X. We also make use of the shorthand $\widehat{\Sigma}_{SS} = \frac{1}{n} X_S^T X_S$ and $\widehat{\Sigma}_{S^cS} = \frac{1}{n} X_{S^c}^T X_S$ as well as $\Pi_S = X_S(\widehat{\Sigma}_{SS})^{-1} X_S^T$ to denote the projector on the range of X_S.

At the core of our constructive procedure is the following convex-analytic result, which characterizes an optimal primal-dual pair for which the primal solution $\widehat{B}$ correctly recovers the support set S:

Lemma 2. *Suppose that there exists a primal-dual pair $(\widehat{B}, \widehat{Z})$ that satisfies the conditions:*

$$\widehat{Z}_S = \zeta(\widehat{B}_S) \tag{28a}$$

$$\widehat{\Sigma}_{SS}(\widehat{B}_S - B^*_S) - \frac{1}{n} X_S^T W = -\lambda_n \widehat{Z}_S \tag{28b}$$

$$\lambda_n \left\| \widehat{Z}_{S^c} \right\|_{\ell_\infty/\ell_2} := \left\| \widehat{\Sigma}_{S^cS}(\widehat{B}_S - B^*_S) - \frac{1}{n} X_{S^c}^T W \right\|_{\ell_\infty/\ell_2} < \lambda_n \tag{28c}$$

$$\widehat{B}_{S^c} = 0. \tag{28d}$$

Then $(\widehat{B}, \widehat{Z})$ is a primal-dual optimal solution to the block-regularized problem, with $\widehat{S}(\widehat{B}) = S$ by construction. If $\widehat{\Sigma}_{SS} \succ 0$, then $\widehat{B}$ is the unique optimal primal solution.

See Appendix A for the proof of this claim. Based on Lemma 2, we proceed to construct the required primal dual pair $(\widehat{B}, \widehat{Z})$ as follows. First, we set $\widehat{B}_{S^c} = 0$, so that condition (28d) is satisfied. Next, we specify the pair $(\widehat{B}_S, \widehat{Z}_S)$ by solving the following restricted version of the SOCP:

$$\widehat{B}_S = \arg \min_{B_S \in \mathbb{R}^{s \times K}} \left\{ \frac{1}{2n} \left\| \left\| Y - X \begin{bmatrix} B_S \\ 0_{S^c} \end{bmatrix} \right\| \right\|_F^2 + \lambda_n \|B_S\|_{\ell_1/\ell_2} \right\}. \tag{29}$$

Since $s < n$, the empirical covariance (sub)matrix $\widehat{\Sigma}_{SS} = \frac{1}{n} X_S^T X_S$ is strictly positive definite with probability one, which implies that the restricted problem (29) is strictly convex and therefore has a unique optimum $\widehat{B}_S$. We then choose $\widehat{Z}_S$ to be the solution of equation (28b). Since any such matrix $\widehat{Z}_S$ is also a dual solution to the SOCP (29), it must be an element of the subdifferential $\partial \left\| \widehat{B}_S \right\|_{\ell_1/\ell_2}$.

It remains to show that this construction satisfies conditions (28a) and (28c). In order to satisfy condition (28a), it suffices to show that no row of the solution $\widehat{B}_S$ is identically zero. From equation (28b) and using the invertibility of the empirical covariance matrix $\widehat{\Sigma}_{SS}$, we may solve as follows

$$(\widehat{B}_S - B^*_S) = \left(\widehat{\Sigma}_{SS} \right)^{-1} \left[\frac{X_S^T W}{n} - \lambda_n \widehat{Z}_S \right] =: U_S. \tag{30}$$

Note that for any row $i \in S$, by the triangle inequality, we have

$$\|\widehat{\beta}_i\|_2 \geq \|\beta^*_i\|_2 - \|U_S\|_{\ell_\infty/\ell_2}.$$

Therefore, in order to show that no row of $\widehat{B}_S$ is identically zero, it suffices to show that the event

$$\mathcal{E}(U_S) := \left\{ \|U_S\|_{\ell_\infty/\ell_2} \leq \frac{1}{2} b^*_{\min} \right\} \tag{31}$$

occurs with high probability. (Recall from equation (13) that the parameter $b^*_{\min}$ measures the minimum ℓ_2-norm of any row of B^*_S.) We establish this result in Section 3.2.

Turning to condition (28c), by substituting expression (30) for the difference $(\widehat{B}_S - B^*_S)$ into equation (28c), we obtain a $(p-s) \times K$ random matrix V_{S^c}, with rows indexed by S^c. For any index $j \in S^c$, the corresponding row vector $V_j \in \mathbb{R}^K$ is given by

$$V_j \;:=\; X_j^T \left([\Pi_S - I_n] \frac{W}{n} - \lambda_n \frac{X_S}{n} (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S \right). \tag{32}$$

In order for condition (28c) to hold, it is necessary and sufficient that the probability of the event

$$\mathcal{E}(V_{S^c}) \;:=\; \left\{ \|V_{S^c}\|_{\ell_\infty/\ell_2} \,<\, \lambda_n \right\} \tag{33}$$

converges to one as n tends to infinity. Consequently, the remainder (and bulk) of the proof is devoted to showing that the probabilities $\mathbb{P}[\mathcal{E}(U_S)]$ and $\mathbb{P}[\mathcal{E}(V_{S^c})]$ both converge to one under the specified conditions.

3.2 Analysis of $\mathcal{E}(U_S)$: Correct inclusion of supporting covariates

This section is devoted to the analysis of the event $\mathcal{E}(U_S)$ from equation (31), and in particular showing that its probability converges to one under the specified scaling. We begin by defining

$$\widetilde{W} \;:=\; \frac{1}{\sqrt{n}} (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} X_S^T W.$$

With this notation, we have

$$U_S \;=\; \widehat{\Sigma}_{SS}^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} - \lambda_n (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S.$$

Using this representation and the triangle inequality, we have

$$\begin{aligned} \|U_S\|_{\ell_\infty/\ell_2} &\leq \left\| (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} \right\|_{\ell_\infty/\ell_2} + \lambda_n \left\| (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S \right\|_{\ell_\infty/\ell_2} \\ &\leq \underbrace{\left\| (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} \right\|_{\ell_\infty/\ell_2}}_{T_1} + \underbrace{\lambda_n \left\| (\widehat{\Sigma}_{SS})^{-1} \right\|_\infty}_{T_2}, \end{aligned}$$

where the form of T_2 in the second line uses a standard matrix norm bound (see equation (42a) in Appendix B), and the fact that $\left\| \widehat{Z}_S \right\|_{\ell_\infty/\ell_2} \leq 1$.

Using the triangle inequality, we bound T_2 as follows:

$$\begin{aligned} T_2 &\leq \lambda_n \left\{ \left\| (\Sigma_{SS})^{-1} \right\|_\infty + \left\| (\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1} \right\|_\infty \right\} \\ &\leq \lambda_n \left\{ D_{\max} + \sqrt{s} \left\| (\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1} \right\|_2 \right\} \\ &\leq \lambda_n \left\{ D_{\max} + \sqrt{s} \left\| (\Sigma_{SS})^{-1} \right\|_2 \left\| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s \right\|_2 \right\} \\ &\leq \lambda_n \left\{ D_{\max} + \frac{\sqrt{s}}{C_{\min}} \left\| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s \right\|_2 \right\}, \end{aligned}$$

which defines $\widetilde{X}_S$ as a random matrix with i.i.d. standard Gaussian entries. From concentration results in random matrix theory (see appendix C), for $s/n \to 0$, we have

$$\left\| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s \right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{s}{n}}\right)$$

with probability $1 - \exp(-\Theta(n))$. Overall, we conclude that

$$T_2 \leq \lambda_n \left\{ D_{\max} + \mathcal{O}\left(\sqrt{\frac{s^2}{n}}\right) \right\}$$

with probability $1 - \exp(-\Theta(n))$.

Turning now to T_1, note that conditioned on X_S, we have $(\mathrm{vec}(\widetilde{W}) \mid X_S) \sim N(\vec{0}_{s\times K}, I_s \otimes I_K)$ where $\mathrm{vec}(A)$ denotes the vectorization of matrix A . Using this fact and the definition of the block ℓ_∞/ℓ_2 norm,

$$\begin{aligned} T_1 &= \max_{i\in S} \left\| e_i^T (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} \right\|_2 \\ &\leq \left\| (\widehat{\Sigma}_{SS})^{-1} \right\|_2^{1/2} \left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \right]^{1/2}, \end{aligned}$$

which defines ζ_i^2 as independent χ^2-variates with K degrees of freedom. Using the tail bound in Lemma 8 (see Appendix F) with $t = 2K \log s > K$, we have

$$\mathbb{P}\left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \geq \frac{4K \log s}{n} \right] \leq \exp\left(-2K \log s \left(1 - 2\sqrt{\frac{1}{2\log s}} \right) \right) \to 0.$$

Defining the event $\mathcal{T} := \left\{ \left\| (\widehat{\Sigma}_{SS})^{-1} \right\|_2 \leq \frac{2}{C_{\min}} \right\}$, we have $\mathbb{P}[\mathcal{T}] \geq 1 - 2\exp(-\Theta(n))$, again using concentration results from random matrix theory (see Appendix C). Therefore,

$$\begin{aligned} \mathbb{P}\left[T_1 \geq \sqrt{\frac{8K \log s}{C_{\min} n}} \right] &\leq \mathbb{P}\left[T_1 \geq \sqrt{\frac{8K \log s}{C_{\min} n}} \,\middle|\, \mathcal{T} \right] + \mathbb{P}[\mathcal{T}^c] \\ &\leq \mathbb{P}\left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \geq \frac{4K \log s}{n} \right] + 2\exp(-\Theta(n)) \\ &= \mathcal{O}\left(\exp(-\Theta(\log s)) \right) \to 0. \end{aligned}$$

Finally, combining the pieces, we conclude that with probability $1 - \exp(-\Theta(\log s))$, we have

$$\begin{aligned}\frac{\|U_S\|_{\ell_\infty/\ell_2}}{b^*_{\min}} &\leq \frac{1}{b^*_{\min}}[T_1 + T_2] \\ &\leq \frac{1}{b^*_{\min}}\left[\mathcal{O}\left(\sqrt{\frac{\log s}{n}}\right) + \lambda_n\left(D_{\max} + \mathcal{O}\left(\sqrt{\frac{s^2}{n}}\right)\right)\right]\end{aligned}$$

With the assumed scaling $n = \Omega(s\log(p-s))$, we have

$$\frac{\|U_S\|_{\ell_\infty/\ell_2}}{b^*_{\min}} \leq \frac{1}{b^*_{\min}}\left[\mathcal{O}\left(\frac{1}{\sqrt{s}}\right) + \lambda_n\left(1 + \mathcal{O}\left(\sqrt{\frac{s}{\log(p-s)}}\right)\right)\right], \tag{34}$$

with probability greater than $1 - 2\exp(-c\log(s)) \to 1$ so that the conditions of Theorem 1 are sufficient to ensure that event $\mathcal{E}(U_S)$ holds with high probability as claimed.

3.3 Analysis of $\mathcal{E}(V_{S^c})$: Correct exclusion of non-support

For simplicity, in the following arguments, we drop the index S^c and write V for V_{S^c}. In order to show that $\|V\|_{\ell_\infty/\ell_2} < \lambda_n$ with probability converging to one, we make use of the decomposition $\frac{1}{\lambda_n}\|V\|_{\ell_\infty/\ell_2} \leq \sum_{i=1}^3 T_i'$ where

$$T_1' := \frac{1}{\lambda_n}\|\mathbb{E}[V \mid X_S]\|_{\ell_\infty/\ell_2} \tag{35a}$$

$$T_2' := \frac{1}{\lambda_n}\|\mathbb{E}[V|X_S, W] - \mathbb{E}[V|X_S]\|_{\ell_\infty/\ell_2} \tag{35b}$$

$$T_3' := \frac{1}{\lambda_n}\|V - \mathbb{E}[V|X_S, W]\|_{\ell_\infty/\ell_2}. \tag{35c}$$

We deal with each of these three terms in turn, showing that with high probability under the specified scaling of (n, p, s), we have $T_1' \leq (1-\gamma)$, and $T_2' = o_p(1)$, and $T_3' < \gamma$, which suffices to show that $\frac{1}{\lambda_n}\|V\|_{\ell_\infty/\ell_2} < 1$ with high probability.

The following lemma is useful in the analysis:

Lemma 3. *Define the matrix $\Delta \in \mathbb{R}^{s\times K}$ with rows $\Delta_i := U_i/\|\beta_i^*\|_2$. As long as $\|\Delta_i\|_2 \leq 1/2$ for all row indices $i \in S$, we have*

$$\left\|\widehat{Z}_S - \zeta(B_S^*)\right\|_{\ell_\infty/\ell_2} \leq 4\|\Delta\|_{\ell_\infty/\ell_2}.$$

See Appendix D for the proof of this claim.

3.3.1 Analysis of T_1'

Note that by definition of the regression model (4), we have the conditional independence relations

$$W \perp\!\!\!\perp X_{S^c} \mid X_S, \qquad \widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid X_S, \quad \text{and} \quad \widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid \{X_S, W\}.$$

Using the two first conditional independencies, we have

$$\mathbb{E}\left[V \mid X_S\right] \;=\; \mathbb{E}\left[X_{S^c}^T | X_S\right] \left([\Pi_S - I_n] \frac{\mathbb{E}\left[W | X_S\right]}{n} - \lambda_n \frac{X_S}{n} \left(\widehat{\Sigma}_{SS}\right)^{-1} \mathbb{E}\left[\widehat{Z}_S | X_S\right] \right).$$

Since $\mathbb{E}\left[W | X_S\right] = 0$, the first term vanishes, and using $\mathbb{E}\left[X_{S^c}^T | X_S\right] = \Sigma_{S^cS}\Sigma_{SS}^{-1} X_S^T$, we obtain

$$\mathbb{E}\left[V \mid X_S\right] = \lambda_n \Sigma_{S^cS}\Sigma_{SS}^{-1} \mathbb{E}\left[\widehat{Z}_S | X_S\right]. \tag{36}$$

Using the matrix-norm inequality (42a) of Appendix B and then Jensen's inequality yields

$$\begin{aligned} T_1' &= \|\Sigma_{S^cS}\Sigma_{SS}^{-1} \mathbb{E}\left[Z_S | X_S\right]\|_{\ell_\infty/\ell_2} \\ &\leq \left|\!\left|\!\left| \Sigma_{S^cS}\Sigma_{SS}^{-1} \right|\!\right|\!\right|_\infty \mathbb{E}\left[\|Z_S\|_{\ell_\infty/\ell_2} | X_S\right] \\ &\leq (1-\gamma). \end{aligned}$$

3.3.2 Analysis of T_2'

Appealing to the conditional independence relationship $\widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid \{X_S, W\}$, we have

$$\mathbb{E}\left[V \mid X_S, W\right] \;=\; \mathbb{E}\left[X_{S^c}^T \mid X_S, W\right] \left([\Pi_S - I_n] \frac{W}{n} - \lambda_n \frac{X_S}{n} \left(\widehat{\Sigma}_{SS}\right)^{-1} \mathbb{E}\left[\widehat{Z}_S | X_S, W\right] \right).$$

Observe that $\mathbb{E}\left[\widehat{Z}_S | X_S, W\right] = \widehat{Z}_S$ because (X_S, W) uniquely specifies $\widehat{B}_S$ through the convex program (29), and the triple $(X_S, W, \widehat{B}_S)$ defines $\widehat{Z}_S$ through equation (28b). Moreover, the noise term disappears because the kernel of the orthogonal projection matrix $(I_n - \Pi_S)$ is the same as the range space of X_S, and

$$\begin{aligned} \mathbb{E}\left[X_{S^c}^T \mid X_S, W\right][\Pi_S - I_n] &= \mathbb{E}\left[X_{S^c}^T \mid X_S\right][\Pi_S - I_n] \\ &= \Sigma_{S^cS}\Sigma_{SS}^{-1} X_S^T [\Pi_S - I_n] \;=\; 0. \end{aligned}$$

We have thus shown that $\mathbb{E}\left[V \mid X_S, W\right] = -\frac{\lambda_n}{n}\Sigma_{S^cS}\Sigma_{SS}^{-1}\widehat{Z}_S$, so that we can conclude that

$$\begin{aligned} T_2' &\leq \left|\!\left|\!\left| \Sigma_{S^cS}(\Sigma_{SS})^{-1} \right|\!\right|\!\right|_\infty \|\widehat{Z}_S - \mathbb{E}\left[\widehat{Z}_S | X_S\right]\|_{\ell_\infty/\ell_2} \\ &\leq (1-\gamma)\, \mathbb{E}\left[\left\| \widehat{Z}_S - Z_S^* \right\|_{\ell_\infty/\ell_2} \right] + (1-\gamma) \left\| \widehat{Z}_S - Z_S^* \right\|_{\ell_\infty/\ell_2} \\ &\leq (1-\gamma)\, 4 \Big\{ \mathbb{E}\left[\|\Delta\|_{\ell_\infty/\ell_2}\right] + \|\Delta\|_{\ell_\infty/\ell_2} \Big\}, \end{aligned}$$

where the final inequality uses Lemma 3. Under the assumptions of Theorem 1, this final term is of order $o_p(1)$, as shown in Section 3.2.

3.3.3 Analysis of T_3'

This third term requires a little more care. We begin by noting that conditionally on X_S and W, each vector $V_j \in \mathbb{R}^K$ is normally distributed. Since $\mathrm{Cov}(X^{(j)} \mid X_S, W) = (\Sigma_{S^c|S})_{jj}\, I_n$, we have

$$\mathrm{Cov}(V_j \mid X_S, W) \;=\; M_n \left(\Sigma_{S^c|S}\right)_{jj}$$

where the $K \times K$ random matrix $M_n = M_n(X_S, W)$ is given by

$$M_n \;\; := \;\; \frac{\lambda_n^2}{n} \widehat{Z}_S^T (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S + \frac{1}{n^2} W^T (\Pi_S - I_n) W. \tag{37}$$

Conditionally on W and X_S, the matrix M_n is fixed, and we have

$$\left(\|V_j - \mathbb{E}\,[V_j \;\mid\; X_S,\; W]\|_2^2 \,|\, W, X_S \right) \;\; \stackrel{d}{=} \;\; \left(\Sigma_{S^c \,|\, S}\right)_{jj} \xi_j^T M_n \xi_j.$$

where $\xi_j \sim N(\vec{0}_K, I_K)$. Since $(\Sigma_{S^c \,|\, S})_{jj} \leq (\Sigma_{S^c S^c})_{jj} \;\leq\; C_{\max}$ for all j, we have

$$\max_{j \in S^c} (\Sigma_{S^c \,|\, S})_{jj} \; \xi_j^T M_n \xi_j \;\; \leq \;\; C_{\max} \; |\!|\!| M_n |\!|\!|_2 \; \max_{j \in S^c} \|\xi_j\|_2^2$$

where $|\!|\!| M_n |\!|\!|_2$ is the spectral norm.

We now state a result that provides control on this spectral norm. Intuitively, this result is based on the fact that the matrix $\frac{n}{\lambda_n^2} M_n$ is a random matrix that concentrates in spectral norm around the matrix $M^* = {Z_S^*}^T (\Sigma_{SS})^{-1} Z_S^*$, where $Z_S^* = \zeta(B_S^*)$, and the fact that the spectral norm of M^* is directly proportional to the defined sparsity/overlap function $\psi(B^*) := |\!|\!| \zeta(B_S^*)^T (\Sigma_{SS})^{-1} \zeta(B_S^*) |\!|\!|_2$.

Lemma 4. *For any $\delta > 0$, define the event*

$$\mathcal{T}(\delta) \;\; := \;\; \Big\{ \; |\!|\!| M_n |\!|\!|_2 \; \leq \lambda_n^2 \; \frac{\psi(B^*)}{n} \; (1 + \delta) \Big\}. \tag{38}$$

Under the conditions of Theorem 1, for any $\delta > 0$, there is some $c_1 > 0$ such that $\mathbb{P}[\mathcal{T}(\delta)^c] \leq 2 \exp(-c_1 \; \log s) \; \to 0$.

See Appendix E for the proof of this lemma.

Using Lemma 4, we can now complete the proof. For any fixed $\delta > 0$ (which can be made arbitrarily small), we have

$$\mathbb{P}[T_3' \geq \gamma] \;\; \leq \;\; \mathbb{P}[T_3' \geq \gamma \;\mid\; \mathcal{T}(\delta)] + \mathbb{P}[\mathcal{T}(\delta)^c].$$

Since $\mathbb{P}[\mathcal{T}(\delta)^c] \to 0$ from Lemma 4, it suffices to deal with the first term. Conditioning on the event $\mathcal{T}(\delta)$, we have

$$\mathbb{P}[T_3' \geq \gamma \;\mid\; \mathcal{T}(\delta)] \;\; \leq \;\; \mathbb{P}\left[\max_{j \in S^c} \|\xi_j\|_2^2 \geq \frac{\gamma^2}{C_{\max}} \; \frac{n}{\psi(B^*) \; (1 + \delta)} . \right]$$

Define the quantity $t^*(n, B^*) \;:=\; \frac{1}{2} \frac{\gamma^2}{C_{\max}} \; \frac{n}{\psi(B^*) \; (1+\delta)}$, and note that $t^* \to +\infty$ under the specified scaling of (n, p, s). By applying Lemma 8 from Appendix F on large deviations for χ^2-variates with $t = t^*(n, B^*)$, we obtain

$$\begin{aligned} \mathbb{P}[T_3' \geq \gamma \;\mid\; \mathcal{T}(\delta)] \;\; &\leq \;\; (p - s) \; \exp\left(-t^* \left[1 - 2 \sqrt{\frac{K}{t^*}} \right] \right) \\ &\leq \;\; (p - s) \; \exp\left(-t^* \, (1 - \delta) \right), \end{aligned} \tag{39}$$

for (n, p, s) sufficiently large. Thus, the bound (39) tends to zero at rate $\mathcal{O}(\exp(-c\log(p-s)))$ as long as there exists $\nu > 0$ such that we have $(1-\delta)\, t^*(n, B^*) > (1+\nu)\log(p-s)$, or equivalently

$$n \quad > \quad (1+\nu)\frac{(1+\delta)}{(1-\delta)}\,\frac{C_{\max}}{\gamma^2}\;\left[2\,\psi(B^*)\;\log(p-s)\right],$$

as claimed.

4 Discussion

In this paper, we have analyzed the high-dimensional behavior of block-regularization for multivariate regression problems, and shown that its behavior is governed by the sample complexity parameter

$$\theta_{\ell_1/\ell_2}(n, p, s) \quad := \quad n/[2\psi(B^*)\log(p-s)],$$

where n is the sample size, p is the ambient dimension, and $\psi(\cdot)$ is a sparsity-overlap function the measures a combination of the sparsity and overlap properties of the true regression matrix B^*.

There are a number of open questions associated with this work. First, note that the current paper provides only an achievability condition (i.e., support recovery can be achieved once the control parameter is larger than some finite critical threshold $t^\star$). However, based both on empirical results (see Figures 2 and 3) and technical aspects of the proof, we conjecture that our characterization is in fact sharp, meaning that the block-regularized convex program (7) fails to recover the support once the control parameter θ_{ℓ_1/ℓ_2} drops below some critical threshold. Indeed, this conjecture is consistent in the special case of univariate regression with $K = 1$, where it is known (Wainwright, 2006) that the Lasso fails once the ratio $n/[2s\log(p-s)]$ falls below a critical threshold. Secondly, the current work applies to the "hard"-sparsity model, in which a subset S of the regressors are non-zero, and the remaining coefficients are zero. As with the ordinary Lasso, it would also be interesting to study block-regularization under soft sparsity models (e.g., ℓ_q "balls" for coefficients, with $q < 1$), under an alternative loss function such as mean-squared error, as opposed to the exact support recovery criterion considered here.

Acknowledgements

This research was partially supported by NSF grants DMS-0605165 and CCF-0545862 to MJW and by NSF Grant 0509559 and DARPA IPTO Contract FA8750-05-2-0249 to MIJ.

A Proof of Lemma 2

Using the notation β_i to denote a row of B and denoting by

$$\mathcal{K} := \{(w, v) \in \mathbb{R}^K \times \mathbb{R} \mid \|w\|_2 \leq v\} \tag{40}$$

the usual second-order cone (SOC), we can rewrite the original convex program (7) as

$$\begin{aligned} \min_{\substack{B \in \mathbb{R}^{p \times K} \\ b \in \mathbb{R}^p}} \quad & \frac{1}{2n} |||Y - XB|||_F^2 + \lambda_n \sum_{i=1}^p b_i \\ \text{s.t.} \quad & (\beta_i, b_i) \in \mathcal{K}, \quad 1 \leq i \leq p. \end{aligned}$$

We now dualize the conic constraints (Boyd and Vandenberghe, 2004), using conic Lagrange multipliers belonging to the dual cone $\mathcal{K}^* = \{(z,t) \in \mathbb{R}^{K+1} | \, z^T \mathbf{w} + vt \geq 0, (\mathbf{w}, v) \in \mathcal{K}\}$. The second-order cone $\mathcal{K}$ is self-dual (Boyd and Vandenberghe, 2004), so that the convex program (41) is equivalent to

$$\begin{aligned} \min_{\substack{B \in \mathbb{R}^{p \times K} \\ b \in \mathbb{R}^p}} \max_{\substack{Z \in \mathbb{R}^{p \times K} \\ t \in \mathbb{R}^p}} \quad & \frac{1}{2n} |||Y - XB|||_F^2 + \lambda_n \sum_{i=1}^p b_i - \lambda_n \sum_{i=1}^p \left(-z_i^T \beta_i + t_i \, b_i \right) \\ \text{s.t.} \quad & (z_i, t_i) \in \mathcal{K}, \quad 1 \leq i \leq p, \end{aligned}$$

where Z is the matrix whose i^{th} row is z_i.

Since the original program is convex and strictly feasible, strong duality holds and any pair of primal $(B^\star, b^\star)$ and dual solutions $(Z^\star, t^\star)$ has to satisfy the Karush-Kuhn-Tucker conditions:

$$\|\beta_i^\star\|_2 \leq b_i^\star, \quad 1 < i < p \tag{41a}$$

$$\|z_i^\star\|_2 \leq t_i^\star, \quad 1 < i < p \tag{41b}$$

$$z_i^{\star T} \beta_i^\star - t_i^\star b_i^\star = 0, \quad 1 < i < p \tag{41c}$$

$$\nabla_B \left[\frac{1}{2n} |||Y - XB|||_F^2 \right] \Big|_{B=B^\star} + \lambda_n Z^\star = 0 \tag{41d}$$

$$\lambda_n (1 - t_i^\star) = 0 \tag{41e}$$

Since equations (41c) and (41e) impose the constraints $t_i^\star = 1$ and $b_i^\star = \|\beta_i^\star\|_2$, a primal-dual solution to this conic program is determined by $(B^\star, Z^\star)$.

Any solution satisfying the conditions in Lemma 2 also satisfies these KKT conditions, since equation (28b) and the definition (28c) are equivalent to equation (41d), and equation (28a) and the combination of conditions (28d) and (28c) imply that the complementary slackness equations (41c) hold for each primal-dual conic pair (β_i, z_i).

Now consider some other primal solution $\widetilde{B}$; when combined with the optimal dual solution $\widehat{Z}$, the pair $(\widetilde{B}, \widehat{Z})$ must satisfy the KKT conditions (Bertsekas, 1995). But since for $j \in S^c$, we have $\|\hat{z}_j\|_2 < 1$, then the complementary slackness condition (41c) implies that for all $j \in S^c$, $\widetilde{\beta}_j = 0$. This fact in turn implies that the primal solution $\widetilde{B}$ must also be a solution to the restricted convex program (29), obtained by only considering the covariates in the set S or equivalently by setting $B_{S^c} = 0_{S^c}$. But since $s < n$ by assumption, the matrix $X_S^T X_S$ is strictly positive definite with probability one, and therefore the restricted convex program (29) has a unique solution $B_S^\star = \widehat{B}_S$. We have thus shown that a solution $(\widehat{B}, \widehat{Z})$ to the program (7) that satisfies the conditions of Lemma 2, if it exists, must be unique.

B Inequalities with block-matrix norms

In general, the two families of matrix norms that we have introduced, $|\!|\!| \cdot |\!|\!|_{p,q}$ and $\|\cdot\|_{\ell_a/\ell_b}$, are distinct, but they coincide in the following useful special case:

Lemma 5. *For $1 \leq p \leq \infty$ and for r defined by $1/r + 1/p = 1$ we have*

$$\|\cdot\|_{\ell_\infty/\ell_p} \quad = \quad |\!|\!| \cdot |\!|\!|_{\infty,r}.$$

Proof. Indeed, if a_i denotes the i^{th} row of A, then

$$\|A\|_{\ell_\infty/\ell_p} = \max_i \|a_i\|_p = \max_i \max_{\|y_i\|_r \leq 1} y_i^T a_i = \max_{\|y\|_r \leq 1} \max_i |y^T a_i| = \max_{\|y\|_r \leq 1} \|Ay\|_\infty = |\!|\!|A|\!|\!|_{\infty,r}.$$

□

We conclude by stating some useful bounds and relations:

Lemma 6. *Consider matrices $A \in \mathbb{R}^{m \times n}$ and $Z \in \mathbb{R}^{n \times \ell}$ and $p, r > 0$ with $\frac{1}{p} + \frac{1}{r} = 1$, we have:*

$$\|AZ\|_{\ell_\infty/\ell_p} \quad = \quad |\!|\!|AZ|\!|\!|_{\infty,r} \leq |\!|\!|A|\!|\!|_{\infty,\infty} |\!|\!|Z|\!|\!|_{\infty,r} = |\!|\!|A|\!|\!|_{\infty,\infty} \|Z\|_{\ell_\infty/\ell_p}. \tag{42a}$$

$$|\!|\!|A|\!|\!|_r \quad \leq \quad |\!|\!|I_m|\!|\!|_{r,\infty} |\!|\!|A|\!|\!|_{\infty,r} = s^{1/r} \|A\|_{\ell_\infty/\ell_p}. \tag{42b}$$

C Some concentration inequalities for random matrices

In this appendix, we state some known concentration inequalities for the extreme eigenvalues of Gaussian random matrices (Davidson and Szarek, 2001). Although these results hold more generally, our interest here is on scalings (n, s) such that $s/n \to 0$.

Lemma 7. *Let $U \in \mathbb{R}^{n \times s}$ be a random matrix from the standard Gaussian ensemble (i.e., $U_{ij} \sim N(0,1)$, i.i.d.). Then*

$$\mathbb{P}\left[\left|\!\left|\!\left| \frac{1}{n} U^T U - I_{s \times s} \right|\!\right|\!\right|_2 \geq \sqrt{\frac{s}{n}}\right] \quad \leq \quad 2\exp(-cn) \;\to\; 0. \tag{43}$$

This result is adapted easily to more general Gaussian ensembles. Letting $X = U\sqrt{\Lambda}$, we obtain an $n \times s$ matrix with i.i.d. rows, $X_i \sim N(0, \Lambda)$. If the covariance matrix Λ has maximum eigenvalue $C_{\max} < +\infty$, then we have

$$|\!|\!|n^{-1} X^T X - \Lambda|\!|\!|_2 \quad = \quad |\!|\!|\sqrt{\Lambda}[n^{-1} U^T U - I]\sqrt{\Lambda}|\!|\!|_2 \;\leq\; C_{\max} |\!|\!|n^{-1}\, U^T U - I|\!|\!|_2 \tag{44}$$

so that the bound (43) immediately yields an analogous bound on different constants.

The final type of bound that we require is on the difference

$$|\!|\!|(X^T X/n)^{-1} - \Lambda^{-1}|\!|\!|_2,$$

assuming that $X^T X$ is invertible. We note that

$$\begin{aligned} \left|\left|\left|(X^T X/n)^{-1} - \Lambda^{-1}\right|\right|\right|_2 &= \left|\left|\left|(X^T X/n)^{-1}[\Lambda - (X^T X/n)]\Lambda^{-1}\right|\right|\right|_2 \\ &\leq \left|\left|\left|(X^T X/n)^{-1}\right|\right|\right|_2 \left|\left|\left|\Lambda - (X^T X/n)\right|\right|\right|_2 \left|\left|\left|\Lambda^{-1}\right|\right|\right|_2. \end{aligned}$$

As long as the eigenvalues of Λ are bounded below by $C_{\min} > 0$, then $\left|\left|\left|\Lambda^{-1}\right|\right|\right|_2 \leq 1/C_{\min}$. Moreover, since $s/n \to 0$, we have (from equation (44)) that $\left|\left|\left|(X^T X/n)^{-1}\right|\right|\right|_2 \leq 2/C_{\min}$ with probability converging to one exponentially in n. Thus, equation (44) implies the desired bound.

D Proof of Lemma 3

From the previous section, the condition $\|\Delta_i\|_2 \leq 1/2$ implies that $\widehat{\beta}_i \neq \vec{0}$ and hence $\widehat{Z}_i = \widehat{\beta}_i / \|\widehat{\beta}_i\|_2$ for all rows $i \in S$. Therefore, using the notation $Z_i^* = \beta_i^* / \|\beta_i^*\|_2$ we have

$$\begin{aligned} \widehat{Z}_i - Z_i^* &= \frac{\widehat{\beta}_i}{\|\widehat{\beta}_i\|_2} - Z_i^* = \frac{Z_i^* + \Delta_i}{\|Z_i^* + \Delta_i\|_2} - Z_i^* \\ &= Z_i^* \left(\frac{1}{\|Z_i^* + \Delta_i\|_2} - 1 \right) + \frac{\Delta_i}{\|Z_i^* + \Delta_i\|_2}. \end{aligned}$$

Note that, for $z \neq 0$, $g(z, \delta) = \frac{1}{\|z+\delta\|_2}$ is differentiable with respect to δ, with gradient $\nabla_\delta g(z, \delta) = -\frac{z+\delta}{2\|z+\delta\|_2^3}$. By the mean-value theorem, there exists $h \in [0, 1]$ such that

$$\frac{1}{\|z+\delta\|_2} - 1 = g(z, \delta) - g(z, 0) = \nabla_\delta g(z, h\delta)^T \delta = -\frac{(z + h\delta)^T \delta}{2\|z + h\delta\|_2^3},$$

which implies that there exists $h_i \in [0, 1]$ such that

$$\begin{aligned} \|\widehat{Z}_i - Z_i^*\|_2 &\leq \|Z_i^*\|_2 \frac{|(Z_i^* + h_i\Delta_i)^T \Delta_i|}{2\|Z_i^* + h_i\Delta_i\|_2^3} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2} \\ &\leq \frac{\|\Delta_i\|_2}{2\|Z_i^* + h_i\Delta_i\|_2^2} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2}. \end{aligned} \tag{45}$$

We note that $\|Z_i^*\|_2 = 1$ and $\|\Delta_i\|_2 \leq \frac{1}{2}$ imply that $\|Z_i^* + h_i\Delta_i\|_2 \geq \frac{1}{2}$. Combined with inequality (45), we obtain $\|\widehat{Z}_i - Z_i^*\|_2 \leq 4\|\Delta_i\|_2$, which proves the lemma.

E Proof of Lemma 4

With $Z_S^* = \zeta(B_S^*)$, define the $K \times K$ random matrix

$$M_n^* := \frac{\lambda_n^2}{n} (Z_S^*)^T (\widehat{\Sigma}_{SS})^{-1} Z_S^* + \frac{1}{n^2} W^T (I_n - \Pi_S) W$$

and note that (using standard results on Wishart matrices (Anderson, 1984))

$$\mathbb{E}\left[M_n^*\right] = \frac{\lambda_n^2}{n-s-1} (Z_S^*)^T (\Sigma_{SS})^{-1} Z_S^* + \sigma^2 \frac{n-s}{n^2} I_K. \tag{46}$$

To bound M_n in spectral norm, we use the triangle inequality:

$$\|\!|M_n|\!\|_2 \;\leq\; \underbrace{\|\!|M_n - M_n^*|\!\|_2}_{A_1} + \underbrace{\|\!|M_n^* - \mathbb{E}\,[M_n^*]|\!\|_2}_{A_2} + \underbrace{\|\!|\mathbb{E}\,[M_n^*]|\!\|_2}_{A_3}. \tag{47}$$

Considering the term A_1 in the decomposition (47), we have:

$$\begin{aligned}
\|\!|M_n^* - M_n|\!\|_2 &= \frac{\lambda_n^2}{n}\left\|Z_S^*\widehat{\Sigma}_{SS}^{-1}Z_S^* - \widehat{Z}_S\widehat{\Sigma}_{SS}^{-1}\widehat{Z}_S\right\|_2 \\
&= \frac{\lambda_n^2}{n}\left\|Z_S^*\widehat{\Sigma}_{SS}^{-1}(Z_S^* - \widehat{Z}_S) + (Z_S^* - \widehat{Z}_S)\widehat{\Sigma}_{SS}^{-1}(Z_S^* + (\widehat{Z}_S - Z_S^*))\right\|_2 \\
&\leq \frac{\lambda_n^2}{n}\left\|\widehat{\Sigma}_{SS}^{-1}\right\|_2\left\|Z_S^* - \widehat{Z}_S\right\|_2\left(2\,\|\!|Z_S^*|\!\|_2 + \left\|Z_S^* - \widehat{Z}_S\right\|_2\right) \qquad (48)
\end{aligned}$$

Using the concentration results on random matrices in Appendix C, we have the bound $\left\|\widehat{\Sigma}_{SS}^{-1}\right\|_2 \leq 2/C_{\min}$ with probability greater than $1 - 2\exp(-cn)$, and we have $\|\!|Z_S^*|\!\|_2 = \mathcal{O}(\sqrt{s})$ by definition. Moreover, from equation (42b) in Lemma 5, we have $\left\|Z_S^* - \widehat{Z}_S\right\|_2 \leq \sqrt{s}\left\|Z_S^* - \widehat{Z}_S\right\|_{\ell_\infty/\ell_2}$. Using the bound (34) and Lemma 3, we have $\left\|Z_S^* - \widehat{Z}_S\right\|_{\ell_\infty/\ell_2} = o(1)$ with probability greater than $1 - 2\exp(-c\log s)$, so that from equation (48), we conclude that

$$A_1 \;=\; \|\!|M_n^* - M_n|\!\|_2 \;=\; o\left(\frac{\lambda_n^2 s}{n}\right) \qquad \text{w.h.p.} \tag{49}$$

Turning to term A_2, we have the upper bound $A_2 \leq T_1^\dagger + T_2^\dagger$, where

$$T_1^\dagger \;=\; \frac{\lambda_n^2}{n}\,\|\!|Z_S^*|\!\|_2^2\,\left\|\frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1}\right\|_2.$$

We have $T_1^\dagger = o\left(\frac{\lambda_n^2 s}{n}\right)$ with probability greater than $1 - 2\exp(-cn)$, since $\|\!|Z_S^*|\!\|_2^2 \leq s$, and $\left\|\frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1}\right\|_2 = o(1)$ with high probability (see Appendix C). Turning to $T_2^\dagger$, we have with probability greater than $1 - 2\exp(-cn)$,

$$T_2^\dagger \;:=\; \frac{1}{n^2}\,\left\|W^T(I_n - \Pi_S)W - \sigma^2(n-s)I_K\right\|_2 \;=\; \mathcal{O}\left(\frac{1}{n}\right) \;=\; o\left(\frac{\lambda_n^2 s}{n}\right),$$

since $\lambda_n^2 s \to +\infty$. Overall, we conclude that

$$A_2 \;=\; \|\!|M_n^* - \mathbb{E}\,[M_n^*]|\!\|_2 \;=\; o\left(\frac{\lambda_n^2 s}{n}\right) \qquad \text{w.h.p.} \tag{50}$$

Finally, turning to $A_3 = \|\!|\mathbb{E}\,[M_n^*]|\!\|_2$, from equation (46), we have

$$\|\!|\mathbb{E}[M_n^*]|\!\|_2 \;\leq\; \frac{\lambda_n^2\psi(B^*)}{n}\frac{n}{n-s-1} \;+ \frac{\sigma^2}{n}\left(1 - \frac{s}{n}\right) = \;(1 + o(1))\left[\frac{\lambda_n^2\psi(B^*)}{n}\right]. \tag{51}$$

Combining bounds (49), (50), and (51) in the decomposition (47), and using the fact that $\psi(B^*) = \Theta(s)$ (see Lemma 1(a)) yields that

$$\|\!|M_n|\!\|_2 \quad \leq \quad (1 + o(1)) \left[\frac{\lambda_n^2 \psi(B^*)}{n} \right]$$

with probability greater than $1 - 2\exp(c \log s)$, which establishes the claim.

F Large deviations for χ^2-variates

Lemma 8. *Let $Z_1, \ldots, Z_m$ be i.i.d. χ^2-variates with d degrees of freedom. Then for all $t > d$, we have*

$$\mathbb{P}[\max_{i=1,\ldots,m} Z_i \geq 2t] \quad \leq \quad m \, \exp\Big(- t \Big[1 - 2\sqrt{\frac{d}{t}} \Big] \Big). \tag{52}$$

Proof. Given a central χ^2-variate X with d degrees of freedom, Laurent and Massart (1998) prove that $\mathbb{P}[X - d \geq 2\sqrt{d\,x} + 2x] \leq \exp(-x)$, or equivalently

$$\mathbb{P}\left[X \geq x + (\sqrt{x} + \sqrt{d})^2 \right] \quad \leq \quad \exp(-x),$$

valid for all $x > 0$. Setting $\sqrt{x} + \sqrt{d} = \sqrt{t}$, we have

$$\begin{aligned} \mathbb{P}[X \geq 2t] \stackrel{(a)}{\leq} \mathbb{P}\left[X \geq (\sqrt{t} - \sqrt{d})^2 + t \right] &\leq \exp(-(\sqrt{t} - \sqrt{d})^2) \\ &\leq \exp(-t + 2\sqrt{td}) \\ &= \exp\left(-t \left[1 - 2\sqrt{\frac{d}{t}} \right] \right), \end{aligned}$$

where inequality (a) follows since $\sqrt{t} \geq \sqrt{d}$ by assumption. Thus, the claim (52) follows by the union bound. □

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.

Argyriou, A., Evgeniou, T., and Pontil, M. (2006). Multi-task feature learning. In *Advances in Neural Information Processing Systems, 18*. MIT Press, Cambridge, MA.

Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. Technical report, INRIA, Département d'Informatique, Ecole Normale Supérieure.

Bach, F., Lanckriet, G., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proc. Int. Conf. Machine Learning (ICML)*. Morgan Kaufmann.

Bertsekas, D. P. (1995). *Nonlinear programming*. Athena Scientific, Belmont, MA.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge, UK.

Buldygin, V. V. and Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI.

Chen, S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61.

Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL.

Donoho, D. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info Theory*, 47(7):2845–2862.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

Laurent, B. and Massart, P. (1998). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338.

Liu, H. and Zhang, J. (2008). On the $\ell_1 - \ell_q$ regularized regression. Technical Report arXiv:0802.1517v1, Carnegie Mellon University.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462.

Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*.

Obozinski, G., Taskar, B., and Jordan, M. (2007). Joint covariate selection for grouped classification. Technical Report 743, Department of Statistics, University of California, Berkeley.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008). SpAM: sparse additive models. Technical Report arXiv:0711.4555v2, Carnegie Mellon University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051.

Turlach, B., Venables, W., and Wright, S. (2005). Simultaneous variable selection. *Technometrics*, 27:349–363.

Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity using using ℓ_1-constrained quadratic programs. Technical Report 709, Department of Statistics, UC Berkeley.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):4967.

Zhang, H., Liu, H., Wu, Y., and Zhu, J. (2008). Variable selection for the multi-category SVM via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2:1149–1167.

Zhao, P., Rocha, G., and Yu, B. (2007). Grouped and hierarchical model selection through composite absolute penalties. Technical Report 703, Statistics Department, University of California, Berkeley.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567.